Survey Designs and Spatio-Temporal Methods for Disease Surveillance

Dissertation Advisor: Professor Marcello Pagano Lauren Brooke Hund

Survey designs and spatio-temporal methods for disease surveillance

Abstract

By improving the precision and accuracy of public health surveillance tools, we can improve cost-efficacy and obtain meaningful information to act upon. In this dissertation, we propose statistical methods for improving public health surveillance research. In Chapter 1, we introduce a pooled testing option for HIV prevalence estimation surveys to increase testing consent rates and subsequently decrease non-response bias. Pooled testing is less certain than individual testing, but, if more people to submit to testing, then it should reduce the potential for non-response bias. In Chapter 2, we illustrate technical issues in the design of neonatal tetanus elimination surveys. We address identifying the target population; using binary classification via lot quality assurance sampling (LQAS); and adjusting the design for the sensitivity of the survey instrument . In Chapter 3, we extend LQAS survey designs for monitoring malnutrition for longitudinal surveillance programs. By combining historical information with data from previous surveys, we detect spikes in malnutrition rates. Using this framework, we detect rises in malnutrition prevalence in longitudinal programs in Kenya and the Sudan. In Chapter 4, we develop a computationally efficient geostatistical disease mapping model that naturally handles model fitting issues due to temporal boundary misalignment by assuming that an underlying continuous risk surface induces spatial correlation between areas. We apply our method to assess socioeconomic trends in breast cancer incidence in Los Angeles between 1990 and 2000. In Chapter 5, we develop a statistical framework for addressing statistical uncertainty associated with denominator interpolation and with temporal misalignment in disease mapping studies. We propose methods for assessing the impact of the un-

certainty in these predictions on health effects analyses. Then, we construct a general framework for spatial misalignment in regression.

Contents

Acknowledgments

First and foremost, I would like to thank my advisors Brent Coull and Marcello Pagano for all of their time and dedication. Over the last four or five years, they have taught me so much about how to conduct meaningful public health research, and I am very grateful for all of their guidance and patience. I would also like to thank Francesca Dominici for serving on my thesis committee and for providing me with excellent advice and support along the way.

Lastly, thank you to my family and friends for their support, specifically Ann, Jim, and Heather Hund. It's been a long road, and I wouldn't have made it without their prodding along the way.

Estimating HIV prevalence from surveys with low individual consent rates: annealing individual and pooled samples

Lauren Hund and Marcello Pagano

Department of Biostatistics

Harvard School of Public Health

1.1 Introduction

HIV prevalence estimates derived from national population-based surveys are often considered the gold standard of HIV prevalence estimation when non-response rates are low (Martin-Herz et al., 2006; Garcia-Calleja et al., 2006; Mishra et al., 2008; Gouws et al., 2008). However, finding and obtaining a blood sample from all individuals surveyed is a considerable, if not almost impossible, challenge. Frequently, migrant or homeless populations are ignored and a large proportion of the sample does not consent to being tested, potentially inducing (unmeasured) bias in the HIV prevalence estimators (Gouws et al., 2008).

In this paper, we discuss a method for promoting increased testing consent rates. Individual reluctance to test may be influenced by several factors, including those related to social stigma associated with HIV and lack of available treatment for testing individuals (Castro and Farmer, 2005; Vermund and Wilson, 2002). While no consensus has been reached on reasons for test refusal or failing to return for test results, fear is a common theme in such studies (Obermeyer and Osborn, 2007), and there is evidence that those who are aware of their positive HIV status are less likely to consent to testing (Reniers and Eaton, 2009). Additionally, the HIV testing protocol is an important factor in gaining test consent (Reniers et al., 2009). The method of asking for consent, specifically convincing survey participants of the importance of their contribution to fighting the HIV epidemic while assuaging concerns about privacy of test results, could be key in improving test consent rates.

One option for estimating prevalence while preserving the nonidentifiability of individuals, at the cost of greater uncertainty, is pooled testing (Gastwirth and Hammick, 1989), where individual samples are combined to form pooled samples. In this paper, we propose a testing protocol that supplements the presumably more informative individual testing with pooled testing. Each sampled individual is asked to provide a blood sample for disease testing, where the investigators (and by choice the individual as well) learn

the disease status of the individual. If the individual rejects this testing option, we ask if he will provide a non-identifiable blood sample which will be combined with other samples in a pooled test and in which case *no one* knows this individual's test result. If the individual does not consent to pooled or individual testing, then he is not tested for the disease, of course.

Ideally, by providing the pooled testing option, we significantly reduce the amount of missingness in the sample. Pooled testing strategies are frequently used in practice (Tu et al., 1995; Litvak et al., 1994; Quinn et al., 2000; Bilder et al., 2010; McMahan et al., 2011), but to our knowledge, have never been combined with individual test results to construct a potentially even better estimator. In this paper, we propose such an estimator and study its analytical properties. In Section 2 of the paper, we discuss testing consent rates in HIV prevalence estimation surveys and give examples of when non-response bias is an issue in such surveys. We propose an estimator in Section 3 and describe its properties; Section 4 includes a simulation study examining small sample properties of this estimator and illustrating the importance of pool size choice in such a survey design. Section 5 suggests additional adjustments to account for non-response of those who consent to neither pooled nor individual testing.

1.2 Missingness in HIV prevalence estimation surveys

Surveys designed to estimate HIV prevalence can have low testing consent rates, and test refusal is potentially associated with risk of HIV infection. Depending on what is driving test refusal in the population, missingness in a sample may induce bias in the estimator of prevalence (Gouws et al., 2008). Reviews of national HIV prevalence surveys have concluded that, while those who refuse testing may have a higher HIV prevalence, bias induced by missingness is usually negligible because response rates are on average sufficiently high (Garcia-Calleja et al., 2006; Mishra et al., 2008). However, the authors make strong assumptions about missingness patterns in the survey and also reference many surveys in which response rates are low enough that it is difficult to believe that bias

in prevalence estimators is negligible. For instance, the HIV testing consent rate is 62.2% in men and 68.2% in women in the most recent national South African survey (Shisana et al., 2005), and consent rates are even lower in the longitudinal HIV surveillance survey in rural KwaZulu Natal, South Africa, described in Tanser et al. (2008).

A taxonomy of the types of patterns of missingness is useful for analysis (Little and Rubin, 2002). When missingness is at random, survey calibration techniques (such as weight-class adjustments, poststratification, and imputation) allow for adjustment of prevalence estimators to remove bias (Lohr, 1999). All such methods depend on the assumption of missing at random, which states that conditional on covariates, the outcome of interest (HIV status) is independent of the missingness mechanism (test refusal). Many studies have shown that HIV test results are not missing completely at random (see Obermeyer and Osborn (2007) and references within); further, assuming missingness is at random is a strong and untestable assumption. When asking individuals to consent to HIV testing, regardless of how much covariate information is available on these individuals, one could reasonably infer that missingness is nonignorable, is associated with disease status, and cannot be completely explained by individual characteristics. For instance, individual covariate information is likely to be unreliable or sparse when dealing with sensitive topics, such as risky sexual behavior, fidelity, or drug use (Tourangeau and Yan, 2007). Sensitive issues such as partaking in risky sexual behavior are of course associated with HIV status, and studies suggest that there are inconsistencies in reporting of sexual behavior in Demographic Health Surveys (DHS) (Curtis and Sutherland, 2004; de Walque, 2007).

Using DHS data from Zambia, one study recently found that models based on observed covariates (i.e. assume missingness is at random) are insufficient to correct for selection bias in HIV prevalence estimation surveys (Bärnighausen et al., 2011). In this study, 28% of men refused testing; the prevalence estimate of HIV in Zambian men increases from 12% (based on measured HIV status alone and imputation) to 21% upon adjusting for unobserved covariates using a Heckman-type selection model. These results

strongly suggest that bias in prevalence estimates can be very severe when missingness depends on unobserved variables.

When missingness is *not* at random, the (heuristically) most conservative range of estimates for HIV prevalence in a sample calculates the lower bound for prevalence by assuming that all non-responders are HIV negative and the upper bound by assuming all non-responders are HIV positive. Such plausibility bounds are obviously very wide when the proportion of non-responders is high but are also arguably the most honest bounds for our certainty regarding the sample prevalence estimates. Specifically, if only a fraction q of the sample responds to the survey, the prevalence of HIV in the sample is $p = qp_R + (1-q)p_{NR}$, where p_R is the sample prevalence in the responders and p_{NR} denote sample prevalence in the non-responders. Since we only know that p_{NR} is between 0 and 1, the lower bound for prevalence in the sample is qp_R and the upper bound is $qp_R + (1-q)$. The width of this interval is $1 - q$, illustrating the importance of maximizing q in the presence of nonignorable missingness.

As an example, consider the 2004 DHS survey in Malawi (National Statistical Office and ORC Macro, 2005). The overall response rate for HIV testing was 70% in women and 63% in men. Of those interviewed by health workers, 22% refused HIV testing; the remainder of the non-response was driven by inability to locate sampled individuals for testing. In the Lilongwe district, the response rate was only 39%, with 49% of subjects refusing HIV testing and the rest unable to be located. The observed prevalence of HIV for the Lilongwe district was 3.7% with 95% CI [sic] (1.0%, 6.4%), whereas the observed prevalence in the rest of the country was 13.2% with 95% CI [sic] (12.3%, 14.2%). The HIV prevalence estimates for Lilongwe were deemed "implausibly low" and prevalence was imputed for everyone in the district of Lilongwe based on demographic information obtained in the household survey. The imputed prevalence for the Lilongwe district was estimated at 10.3 % with 95% CI [sic] (9.3%, 11.3%).

Consider the conservative plausibility bounds mentioned above for the Lilongwe district. There were 500 individuals eligible for HIV testing in the district of Lilongwe, but

only 193 of those eligible consented to HIV testing. Based on this information, we deduce that about seven out of the 193 consenters were HIV positive. If we assume all 307 non-consenters were HIV negative, a lower bound for HIV prevalence is 1.4% with 95% CI (0.4%, 2.4%); likewise, if we assume all 307 non-consenters were HIV positive, an upper bound for HIV prevalence is 62.8% with 95% CI (58.6%, 67.0%). By taking the lower confidence bound when we assume all non-responders are negative and the upper confidence bound when we assume all non-responders are HIV positive, we can obtain the most conservative plausibility bounds at the 95% confidence level. In the Lilongwe case, the heuristic "plausibility bounds" for the prevalence of HIV are (0.4%, 67.0%), which now includes the national prevalence estimate for HIV in Malawi. While no one would ever present such wide plausibility bounds, these extreme bounds show the true amount of certainty we have when we know nothing about non-responders. The Lilongwe example illustrates the dangers of high non-response in an HIV prevalence estimation survey and that everything possible should be done to minimize non-response in HIV prevalence estimation surveys.

1.3 Framework for combining individual and pooled test results

In standard HIV testing surveys, individuals are only asked to consent to an HIV test once. Using a pooled testing option, we offer two opportunities to consent to HIV testing. For those who select the non-identifiable pooled testing option, individual blood samples are pooled with $k-1$ other blood samples ($k > 1$), and only the test result of the pool is known to anyone. We delay discussion about appropriate choice of k to below. Though we anticipate that some will still refuse both individual and pooled HIV testing, the intent is to lower missingness in the sample (and the associated inherent bias in the estimator) by including individuals who refuse individual testing but are willing to provide a sample for pooled testing. We propose a combined individual and pooled testing prevalence estimator, for which privacy is preserved but prevalence can be estimated more accurately

than when using only those willing to submit to individual testing.

In order to construct this estimator, we consider a simple random sample (SRS) survey design in which n individuals are sampled from a population of size N with disease prevalence p. The methodology is straightforward to extend to the stratified or cluster sampling case, insofar as pools are composed within the strata/clusters and a sufficient proportion of the sample consents to pooled testing within each stratum/cluster. Assuming we have a simple random sample of the population, the sample can be partitioned into three separate groups: 1) those who consent to testing for a disease, 2) those who only consent to unidentifiable pooled testing, and 3) those who refuse testing altogether. The prevalence in each of these three groups may differ. We now describe a statistical framework for constructing an estimator of prevalence based on the above partitioning of the sample.

Let $Y = (Y_1, Y_2, Y_3)$ be a random variable classifying individuals by their testing consent choices, $Y \sim Multinom(n, q_1, q_2, q_3)$, where $n = Y_1 + Y_2 + Y_3$. Specifically, Y_1 reflects the number who consent to individual testing; Y_2 reflects the number who do not consent to individual testing but consent to pooled testing; and Y_3 reflects the number who do not consent to test at all. Let $X_1|Y_1 =$ number of HIV positive persons who consent to individual test, $X_2|Y_2 =$ number of HIV positive persons who consent to pooled test, and $X_3|Y_3 =$ number of HIV positive persons who do not consent to test. We model $X_i|Y_i \sim Bin(Y_i, p_i), i = \{1, 2, 3\}$. For notational simplicity, we write X_i instead of $X_i|Y_i$. We assume X_1, X_2, and X_3 are independent. Let p denote the overall prevalence of HIV infection in the population, so

$$p = p_1 q_1 + p_2 q_2 + p_3 q_3.$$

Note that we can never know p_3, and any estimator of p will always be biased unless p_3 is equal to the prevalence in the population that consents to test; q_3 is 0 (everyone consents to test); or we adjust the estimator of prevalence based on some known and identifiable structure on p_3, such as $p_3 = p_2$. However, we can estimate the probability of having HIV given that one consents to test by conditioning on the sample size in the consenters,

$n' = Y_1 + Y_2$ and adjusting q_1 and q_2 appropriately. That is, we define $q_1' = q_1/(q_1 + q_2)$ and $q_2' = q_2/(q_1 + q_2)$ and redefine $(Y_1, Y_2) \sim Multinom(n', q_1', q_2')$ or equivalently $Y_1 \sim Bin(n', q_1')$. Therefore, $p_T = p_1 q_1' + p_2 q_2'$. A natural estimator for p_T is $\hat{p}_T = \hat{p}_1 \hat{q}_1' + \hat{p}_2 \hat{q}_2'$, where $\hat{q}_1' = Y_1/n'$, $\hat{q}_1' = Y_2/n'$, $\hat{p}_1 = X_1/Y_1$, and $\hat{p}_2$ is a consistent estimator of p_2 that has yet to be determined. Note that $\hat{q}_1', \hat{q}_2',$ and $\hat{p}_1$ are consistent estimators of q_1', q_2' and p_1, respectively, as $n' \to \infty$. If p_2 is observed, we can express p_T in terms of the sample quantities as:

$$\hat{p}_T = \frac{Y_1}{n'}\frac{X_1}{Y_1} + \frac{Y_2}{n'}\frac{X_2}{Y_2}.$$

However, because of the desire to preserve anonymity, we do not directly observe X_2, the number of HIV positive individuals in the pooled population. Rather, we observe the number of pools that test positive, Z. It is straightforward to show that, conditional on Y_2, $Z \sim Bin(n_p, p_z)$, where $n_p = Y_2/k$ is the total number of pools, *k* is the pool size, and $p_z = 1 - (1 - p_2)^k$. Define $\hat{p}_z = Z/n_p$. It follows that $p_2 = 1 - (1 - p_z)^{1/k}$. Since $\hat{p}_z \xrightarrow{p} p_z$ as $n_p \to \infty$, a consistent estimator for the prevalence in the pooled-consenting population is the maximum likelihood estimator, $\hat{p}_2 = 1 - (1 - \hat{p}_z)^{1/k}$. Assuming p_2 is bounded away from 0 and 1, we know that $\hat{p}_2$ is an asymptotically unbiased estimator of p_2. Conditional on Y_2, the asymptotic variance of $\sqrt{Y_2}\hat{p}_2$ is $(1 - p_2)^2((1 - p_2)^{-k} - 1)/k$ (Tu et al., 1995). The variance of $\hat{p}_2$ increases as k increases.

We can estimate the population prevalence in those who consent to test, p_T, consistently as:

$$\hat{p}_T = \frac{Y_1}{n'}\frac{X_1}{Y_1} + \frac{Y_2}{n'}\hat{p}_2$$

which we refer to as the combined prevalence estimator. It follows that, as $n' \to \infty$, with q_1, q_2 bounded away from 0, $\hat{p}_T$ is unbiased (see Section 1.7.1 for proof), and the variance of $\hat{p}_T$ has the limiting form (see Section 1.7.2 for proof):

$$n' var(\hat{p}_T) = q_1' p_1(1 - p_1) + q_2' \tfrac{1}{k}(1 - p_2)^2((1 - p_2)^{-k} - 1) + q_1' q_2'(p_1 - p_2)^2.$$

A natural large-sample variance estimator is thus:

$$\hat{var}(\hat{p}_T) = \frac{1}{n'}\left[\tfrac{X_1}{n'}(1 - \tfrac{X_1}{Y_1}) + \tfrac{Y_2}{n'}\tfrac{1}{k}(1 - \hat{p}_2)^2((1 - \hat{p}_2)^{-k} - 1) + \tfrac{Y_1}{n'}\tfrac{Y_2}{n'}(\tfrac{X_1}{Y_1}^2 - \hat{p}_2)^2\right]$$

Further, it can be shown that $(\hat{p}_T - p_T)/\sqrt{var(\hat{p}_T)} \sim N(0,1)$ (Section 1.7.3). Therefore, we can define a $100(1-\alpha)\%$ Wald-type confidence interval for $\hat{p}_T$ as $\hat{p}_T \pm z_{\alpha/2}\sqrt{\hat{var}(\hat{p}_T)}$.

1.4 Properties of the combined estimator

In the remainder of this paper, we consider low, moderate, and high population prevalence settings where individual testing consent rates are low. In the low prevalence setting, we assume the prevalence in the individual testers is 5% and the prevalence in the pooled testers is 10%; in the moderate setting, prevalence in individual testers is 15% and in pooled testers is 20%; and in the high prevalence setting, prevalence in the individual testers is 20% and in the pooled testers is 30%. We assume that the sub-population that consents to individual testing constitutes 60% of the total testing population and the sub-population that will only contribute a sample for pooled testing constitutes 40% of the population. We also constrain pool size to be between 3 and 7. While a smaller pool size will always result in a better estimator, pool size must be sufficiently large to protect the confidentiality of the testers; in our simulation, we assume ethical limitations would never mandate having a pool size larger than 7 and use this as our maximum pool size. These settings are important to keep in mind and are referenced throughout the paper as the low, moderate, and high prevalence settings.

The estimator in which we include pooled testers will almost always provide an improvement (in terms of mean-squared error) over the estimator which only offers individual testing. If we only offer individual testing, an estimate of the prevalence in the population is $\hat{p_1} = X_1/Y_1$. Assuming for now that $q_3 = 0$, the bias in $\hat{p}_1$ is $p_1 - p = q_2(p_1 - p_2)$, which is non-zero when $p_1 \neq p_2$ and $q_2 \neq 0$. However, even if $p_1 \approx p_2$, the estimator using pooled samples will usually have a smaller variance than the estimator which does not incorporate pooled testing as long as a sufficient proportion of the population consents to pooled testing. Since the combined estimator is asymptotically unbiased, the asymptotic

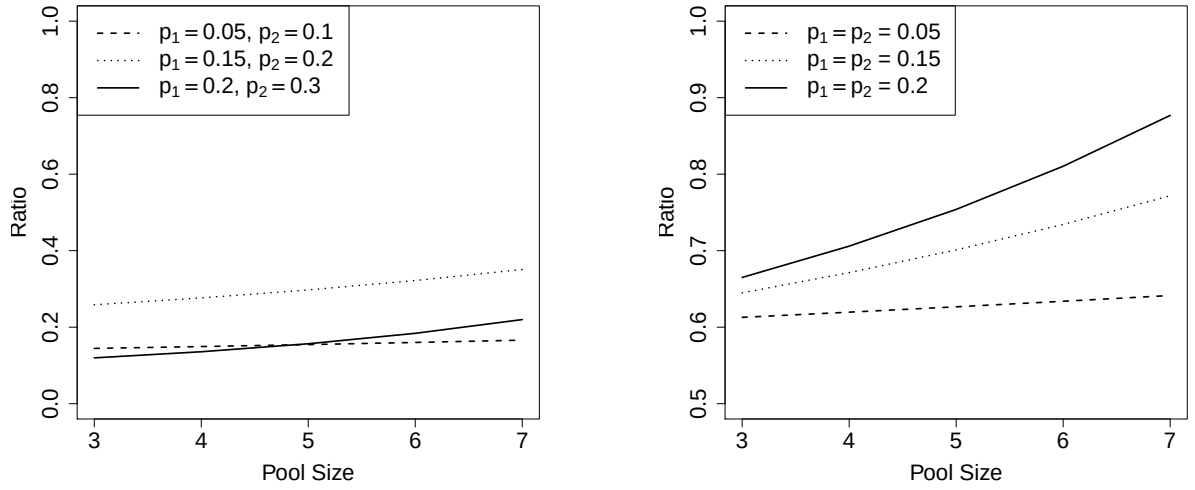
mean-squared error (MSE) of the estimator is:

$$MSE(\hat{p}_T) = \frac{1}{n'}(q_1' p_1(1-p_1) + q_2' \frac{1}{k}(1-p_2)^2((1-p_2)^{-k} - 1) + q_1' q_2' (p_1 - p_2)^2).$$

The estimator using only individual testers has MSE:

$$MSE(\hat{p}_1) = \frac{1}{n' q_1'} p_1(1-p_1) + [q_2(p_1 - p_2)]^2.$$

The ratio of these MSEs is always less than one when pool size is less than 7 for the low, moderate, and high prevalence settings (see Figure 1.1), indicating that the combined estimator outperforms the estimator using only individuals. Indeed, in the situations in which pooled testers have a higher prevalence than individual testers, the MSE ratio ranges between 0.1 and 0.4, and the combined estimator provides substantial improvement over the estimator ignoring pooled testers. Even when the prevalence is the same in the pooled and individual testing populations, the MSE ratio ranges between 0.6 and 0.85, and the combined estimator still outperforms the individuals-only estimator.



(a) (b)

Figure 1.1: Ratio of the asymptotic MSE for the combined estimator to the ratio of the asymptotic MSE for the estimator using only individuals ($\text{MSE}(\hat{p}_T)/\text{MSE}(\hat{p}_1)$) in the low, moderate, and high prevalence settings for two scenarios: (a) pooled testers have a higher prevalence than individual testers, $n' = 1000$; (b) the prevalence in the pooled testers equals that in the individual testers (this ratio is independent of n'). The combined estimator always has lower MSE than the individuals only estimator in these settings.

Only offering pooled testing to everyone in the sample, as suggested in Gastwirth and Hammick (1989), is cheaper than offering an individual and pooled testing option, because fewer tests are performed. For instance, we could design a study which only offers a pooled testing option and estimate prevalence using:

$$\hat{p}_{pool} = 1 - (1 - Z/n_p)^k$$

where Z is the number of positive pools, n_p is the total number of pools, and k is the pool size. Since $\hat{p}_{pool}$ is asymptotically unbiased, the asymptotic MSE of this estimator is

$$MSE(\hat{p}_{pool}) = \tfrac{1}{nk}(1-p)^2[(1-p)^{-k} - 1].$$

Using the ratio $\text{MSE}(\hat{p}_T)/\text{MSE}(\hat{p}_{pool})$, we find that testing using the combined estimator $\hat{p}_T$ results in a smaller asymptotic MSE than the estimator which only offers pooled testing $\hat{p}_{pool}$ (Figure 1.2), assuming the sample size is the same for both estimators. The MSE

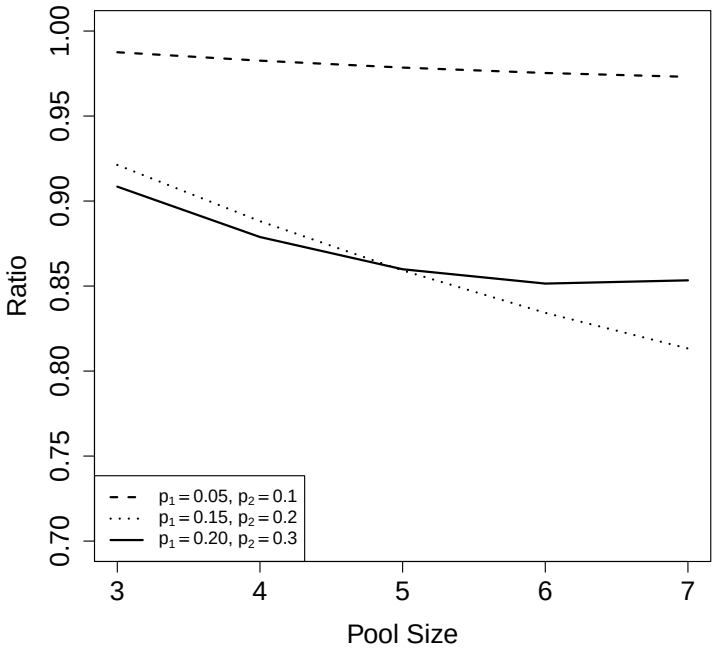


Figure 1.2: Ratio of the MSE for the combined estimator to the ratio of the MSE when everyone is offered pooled testing, $\text{MSE}(\hat{p}_T)/\text{MSE}(\hat{p}_{pool})$, as a function of pool size for the low, moderate, and high prevalence settings when pooled testers have a higher prevalence than individual testers. The combined estimator always has lower MSE than the estimator where everyone is offered pooled testing in these settings.

for the combined estimator is 10% less than the MSE for the pooled testing only estimator in the moderate and high prevalence settings, with less reduction in MSE in the low prevalence setting. The combined estimator provides an improvement in MSE because of the previously mentioned fact that the variance of the pooled prevalence estimator always decreases as the pool size decreases; intuitively, individual test results provide more information than pooled test results on the same number of people, so providing an individual testing option is optimal. Further, if everyone is offered pooled testing, individual results are no longer available to those who are interested in learning their HIV status and thus may be unethical (Diaz et al., 2005). And lastly, the survey protocol we suggest gives individuals two opportunities to consent to testing (pooled or individual), rather than only asking individuals to test once as in the pooled-testing only design, which could help increase consent rates. Therefore, having both pooled and individual testing options is advantageous.

Pooled prevalence estimators are biased in finite samples (Tu et al., 1995), and consequently, $\hat{p}_T$ is only asymptotically unbiased (Section 1.7.4):

$$E(\hat{p}_T) = p_T + \frac{k-1}{2n'(1-p_2)} E\left(Y_2 var(\hat{p}_2)\right) + O\left(\left(\frac{n'}{k}\right)^{-\frac{3}{2}}\right) \neq p_T$$

While replacing an estimator with a jackknifed version of the estimator typically reduces finite sample bias (Quenouille, 1956; Miller, 1974; Shao and Tu, 1995), in simulation, we find that the jackknife estimator provides little improvement over the original estimator (results not shown). Other suggestions for bias correction to the pooled prevalence estimator have been suggested (Hepworth and Watson, 2009). For example, Burrows (1987) suggests the estimator:

$$\tilde{p}_2 = 1 - \left[\frac{2kZ + k - 1}{2kn_p + k - 1}\right]^{1/k}$$

which removes the bias of order n'^{-1}. We can use the Burrows estimator to define a new prevalence estimator $\tilde{p}_T$, which is constructed by substituting $\tilde{p}_2$ for $\hat{p}_2$ in the combined estimator. This new estimator $\tilde{p}_T$ has much smaller finite sample bias than $\hat{p}_T$ in small samples. In Figure 1.3, we plot the percent bias in the prevalence estimator for $\hat{p}_T$ and

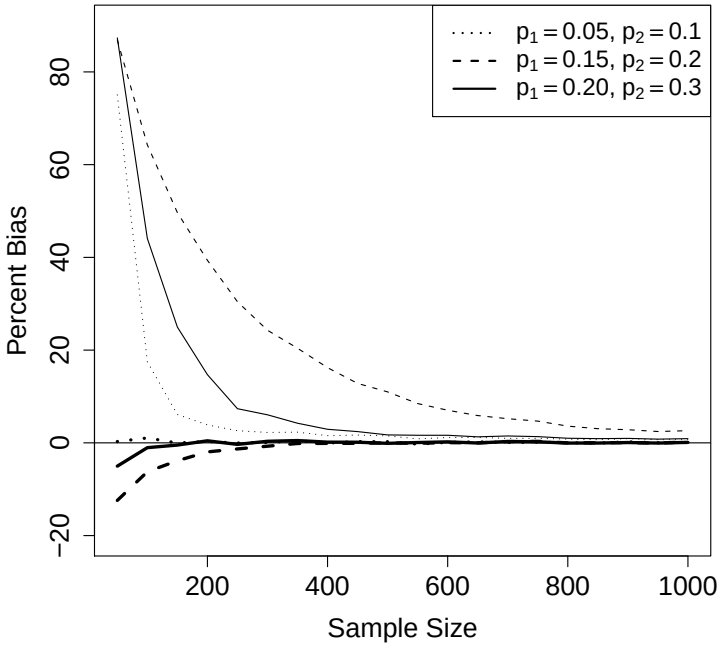


Figure 1.3: Percent bias in the MLE estimator $\hat{p}_T$ (thin lines) and the Burrows estimator $\tilde{p}_T$ (bold lines) for pool size $k = 7$ as a function of sample size for low, moderate, and high prevalence settings. Using the Burrows estimator results in a substantial reduction in finite sample bias.

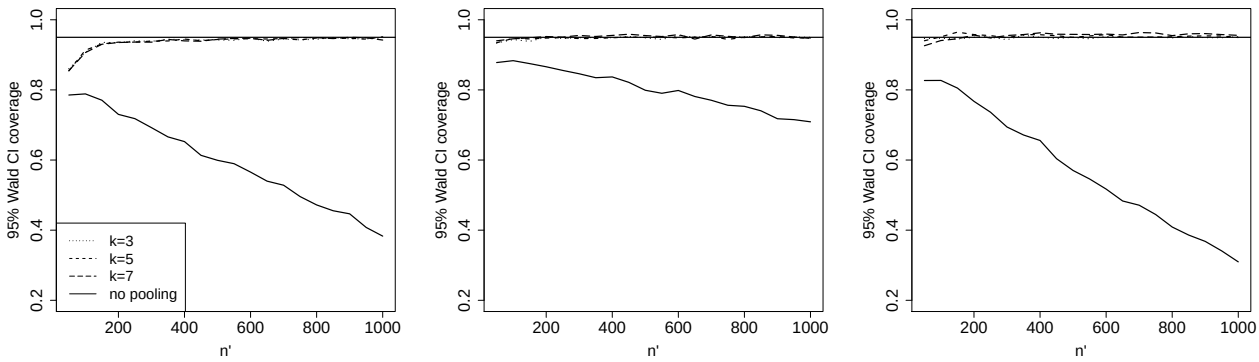
$\tilde{p}_T$ for pool size $k = 7$ (the size for which we see the greatest finite-sample bias). The original estimator $\hat{p}_T$ always overestimates the prevalence, with the severity of the bias decreasing as the sample size increases. The Burrows estimator $\tilde{p}_T$ has negligible bias, even for sample sizes as small as 100. Consequently, we recommend using $\tilde{p}_T$ in practice rather than $\hat{p}_T$.

1.4.1 Simulation study assessing finite sample properties of the combined estimator

Pooling has its limitations that are a function of prevalence. When the prevalence is high, then, to be informative, the pools must be so small as not to have all the pools test positive (Tu et al., 1995; Burrows, 1987). On the other hand, to retain anonymity, the pool sizes cannot be too small. Statistically, pooled estimators are potentially unstable when

the prevalence in the pooled-sample population (p_2) is high or when the number of individuals consenting to pooled testing (Y_2) is small. In the case of most diseases that are not extremely rare, such as HIV, the disease prevalence is typically high enough that some pools will test positive, and we are not concerned with zero pools testing positive. However, in moderate to high prevalence settings, the probability that all pools will test positive must also be addressed. This probability is $P(Z = n_p) = (1 - (1 - p_2)^k)^{n_p}$, which decreases as n_p increases and/or k and p_2 decrease. Therefore, choosing a sufficiently small pool size k and obtaining a sufficiently large number of pools n_p are necessary to ensure that the estimate of the population prevalence in the pooled testing group is reasonable. Note that the lower bound for k is determined by how large the pools should be to assuage concerns about identifiability of test results (see Section 1.6).

In a simulation study, we evaluate maximum pool sizes and minimum number of pools such that the bias and standard error of $\tilde{p}_T$ are small and the 95% Wald confidence interval coverage of $\tilde{p}_T$ is near 0.95. Individuals who do not consent to testing at all are ignored throughout the simulations. Simulation parameters are chosen to reflect low, moderate, and high prevalence settings which have low testing consent rates for individuals as described in Section 1.4. We perform the simulation study for pool sizes 3, 5, and



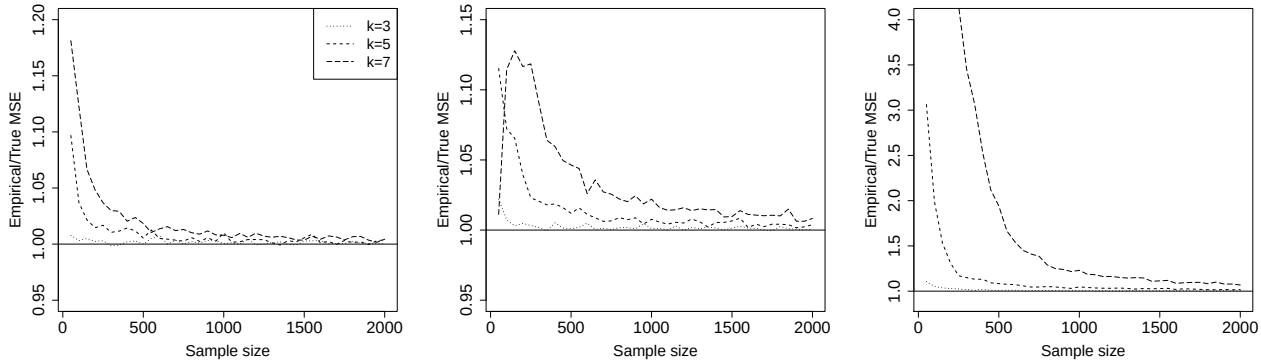
(a) (b) (c)

Figure 1.4: 95% confidence interval coverage for $\tilde{p}_T$ as a function of sample size calculated using various pool sizes in the (a) low, (b) moderate, and (c) high prevalence setting as a function of the sample size.

7 (with 5,000 iterations each). Wald 95% confidence interval coverage is shown in Figure 1.4.

The 95% Wald confidence interval performs well for the combined estimator, with coverage lingering around 95% for moderate sample sizes. The confidence interval coverage drops below 60% very quickly when the pooled testers are ignored. As in the Lilongwe example, confidence intervals are misleading when selection bias exists in the sample.

In small sample sizes for the moderate and high prevalence settings, the empirical standard error for the combined estimator is much larger than the derived standard error (results not shown), due to the fact that all of the pools test positive in a substantial proportion of the simulation runs. The derived large-sample standard error is not valid when all pools test positive, and, in such settings, using the pooled prevalence estimator in practice is not advised. Further, finite sample bias is problematic in small sample sizes when prevalence is moderate to high. Before using the asymptotic normality and variance formula for the combined estimator, it is important to know how many pooled testers are required for these asymptotics to be valid. In order to assess when the large-sample asymptotics hold and the combined prevalence estimator is valid, we calculate



(a) (b) (c)

Figure 1.5: Plot of the ratio of the empirical to the true MSE of the combined estimator as a function of sample size for the (a) low, (b) moderate, and (c) high prevalence settings. When asymptotic results are valid, this ratio will be close to 1.

Table 1.1: Sample size (number of pooled testers) required to have an empirical MSE: true MSE ratio < 1.05.

Pool size	Low Prevalence	Moderate Prevalence	High Prevalence
3	50 (20)	50 (20)	150 (60)
5	100(40)	200 (80)	700 (280)
7	200(80)	500 (200)	> 2000 (> 800)

the ratio of the empirical MSE and the asymptotic MSE (Figure 1.5). The asymptotic MSE is described above, and the empirical MSE is defined as the square of the average empirical bias in the combined estimator added to the empirical variance of the combined estimator in the 5000 simulations. Since both empirical variance and bias should be higher than the asymptotic variance and bias in finite samples, this ratio should provide a good metric for gauging the validity of our estimator. When this ratio is less than 1.05, we declare the estimator to be valid.

Table 1.1 provides suggestions as to minimum sample size and pool size required in the low, moderate, and high prevalence settings in order to obtain a valid estimator. We recommend not using pool sizes over 5 (preferably 3) in the high prevalence setting.

1.5 Adjusting for individuals who refuse testing

Ideally, in a disease prevalence estimation survey, all sampled individuals will consent to test, either as an individual or in a pool. However, in practice, we anticipate that a certain proportion of the population, q_3, will refuse testing altogether. Unless we can assume test status is missing completely at random, accounting for this missingness induced by test refusal is key to constructing an unbiased estimator of prevalence. As previously discussed, assuming data is missing at random may be a poor assumption in such settings. A more reasonable assumption might be that that those who refuse testing are more similar to those who consent to pooled testing than they are to those who consent to individual testing. With this motivation, we propose a weight-class adjustment (also called response

propensity weighting) to the estimator to improve the precision of population prevalence estimates (Lohr, 1999).

In order to adjust the prevalence estimator, we divide the sample of size n into j different strata, $j = 1, ..., J$. Denote the number of individuals sampled in the j^{th} stratum as n_j, and assume that n'_j individuals in stratum j consent to testing. If we have obtained a simple random sample of the population, a naive estimator for prevalence, without taking into account nonconsenters is:

$$\hat{p} = \sum_j \frac{n'_j}{\sum_j n'_j} \hat{p}_j,$$

where $\hat{p}_j$ is the prevalence estimator in stratum j. This estimator is equivalent to estimating prevalence by calculating the number of disease positive individuals in a sample who consent to testing and dividing by the total number of consenters. Hence, this estimator relies on the assumption that we obtained a representative sample of the population, namely that $n'_j / \sum_j n'_j = n_j / \sum_j n_j$.

In order to adjust for non-consenters, we can weight each consenting individual in the sample by the inverse probability that they consent to testing. This method of propensity score weighting produces an unbiased estimator of prevalence when consenters and non-consenters within stratum j are alike with respect to HIV status (that is, there are no unmeasured confounders within stratum j). Using propensity score weighting, the adjusted prevalence estimate becomes:

$$\hat{p} = \sum_j \frac{n_j}{n} \hat{p}_j.$$

Propensity weighting adjustments have been discussed frequently in the literature and have disadvantages including inflating the variance when the weights are large (Little, 1986; Little and Vartivarian, 2003; Little, 1988). Such a situation would occur when individuals in a given stratum are very unlikely to participate in a survey. Collapsing strata can be effective in reducing the impact of sparse data and large weights within a stratum if such a situation occurs. Note that rather than dividing the data into strata, propensity

scores can be calculated using logistic regression and weights can be constructed based on predicted probabilities from a logistic regression, as employed in Mishra et al. (2008).

This propensity weighting framework extends naturally to the combined prevalence estimator, assuming that we can construct homogeneous pools based on the j strata. Construction of homogeneous pools is the primary challenge of implementing the weight-class adjustment correction. Choosing appropriate strata requires balancing the need for a sufficient number of pooled testers within each stratum to maintain confidentiality and obtain valid prevalence estimates as well as the need to incorporate a sufficient amount of information about the testers versus non-testers. Assuming we can construct such strata, we can use the weight class adjustment in two different ways: 1) weight everyone in the sample who consents to test by the inverse probability of testing within their respective stratum, or 2) weight only the pooled testers by the inverse probability of testing as a pooled tester, conditional on not testing as an individual. The first method of weight class adjustment assumes that non-testers are similar to testers (pooled or individual) within strata with respect to HIV status, whereas the second method assumes that non-testers are similar to pooled testers within strata. To choose the appropriate adjustment method, reasons for not consenting to test should be obtained from the sample when possible. For instance, if most people will not test because they dislike having blood drawn, then the first method might be more plausible. If hesitance of the pooled testers and non-testers is caused by suspicion of HIV positive status, the second method is more reasonable.

Simpler estimators could also be proposed without employing a weight-class adjustment, which may be more feasible in practice. For instance, one could assume the prevalence of HIV in the non-testers is equal to the prevalence within the pooled testing population and suggest $\hat{p} = \hat{p}_1\hat{q}_1 + \hat{p}_2(\hat{q}_2 + \hat{q}_3)$, which is potentially a better estimator than $\hat{p}_T$ for the prevalence in the population. Lastly, we could assume a linear trend exists between $p_1, p_2,$ and $p_3,$ and define a prevalence estimator as $\hat{p} = \hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2 + \hat{p}_3\hat{q}_3$ using linear extrapolation (*e.g.* $\hat{p}_i = a + bi$). These estimators need to be tested in practice before we can contrast their merits.

1.6 Discussion

When investigators designing a disease prevalence survey anticipate high refusal rates for individual testing due to disease stigma, offering a pooled testing option and combining pooled and individual sample results has the potential to significantly improve precision of prevalence estimates. Further, acquiring blood samples for pooled testing also allows the investigator to compare the prevalence in the individual testing population (p_1) with the prevalence in the pooled testing population (p_2). A test of the hypothesis that $p_1 = p_2$ is simple to construct. This hypothesis test and a corresponding 95% CI for $(\hat{p}_1 - \hat{p}_2)$ can help determine the extent of selection bias in the sample. Evidence that the consenting and part of the refusing populations are not different with respect to disease status is valuable for generalizability of results to the entire population. Note that this is an association test which does not take any covariates into account, though the test could be conducted within strata if sample sizes are large enough.

Techniques have also been developed for regression analyses of disease status on covariates when blood samples are pooled (Vansteelandt et al., 2000; Xie, 2001; Bilder and Tebbs, 2009; Chen et al., 2009). These ideas could easily be extended to the situation in which we have both individual and pooled test results by allowing the pool size to vary in the regression analysis (that is, pool size is 1 or k). Though we do not want to identify individuals within pooled samples, constructing pools that are homogeneous with respect to the covariates of interest increases the precision of the regression coefficient estimates (Vansteelandt et al., 2000).

Our proposed estimator above assumes a perfect test (sensitivity and specificity are equal to one), but extending the estimator to imperfect tests is straightforward, as shown in Tu et al. (1995). Let ϕ and ψ represent test sensitivity and specificity, respectively. The probability that an individual consenter tests positive is $p_1\phi + (1 - p_1)(1 - \psi)$; assuming sensitivity and specificity are the same for pools as for individual tests, the probability that a pool tests positive is $(1 - (1 - p_2)^k)\phi + (1 - p_2)^k(1 - \psi)$. Note that we also make the

relatively mild assumption that $\phi+\psi-1>0$. It follows that $\hat{p}_{1,\phi,\psi}=(X_1/Y_1+1-\psi)/(\phi+\psi-1)$ and $Var(\hat{p}_{1,\phi,\psi})=Var(\hat{p}_1)/(\phi+\psi-1)^2$. Define $\tilde{p}_z$ as Z/n_p when using the standard pooled prevalence estimator; and as $(Z+(k-1)/2k)/(np+(k-1)/2k)$ when the Burrows correction is used. In the pooled setting,

$$\hat{p}_{2,\phi,\psi}=1-\left(\frac{\phi-\tilde{p}_z}{\phi+\psi-1}\right)^{1/k}$$

when $1-\psi\leq\tilde{p}_z\leq\phi$; $\hat{p}_{2,\phi,\psi}=0$ when $0\leq\tilde{p}_z\leq1-\psi$; and $\hat{p}_{2,\phi,\psi}=1$ when $\phi\leq\tilde{p}_z\leq1$. Also, asymptotic normality for $\hat{p}_{2,\phi,\psi}$ holds, where $Var(\hat{p}_{2,\phi,\psi})=Var(\hat{p}_2)/(\phi+\psi-1)^2$. Therefore, when the sensitivity and specificity of a test are known, they are easily incorporated into the framework of the individual and pooled testing prevalence estimator, as $\hat{p}_{T,\phi,\psi}=q_1\hat{p}_{1,\phi,\psi}+q_2\hat{p}_{2,\phi,\psi}$ and $(\hat{p}_{T,\phi,\psi}-p_T)/(\hat{Var}_{\phi,\psi}(\hat{p}_{T,\phi,\psi}))^{1/2}\sim N(0,1)$, where $Var_{\phi,\psi}(\hat{p}_{T,\phi,\psi})$ is simple to calculate by using the same form of the variance as $\hat{p}_T$, but substituting $V_1/(\phi+\psi-1)^2, V_2/(\phi+\psi-1)^2$ for V_1, V_2 (see Section 1.7.3). Sample variance is calculated by substituting $\hat{p}_{1,\phi,\psi}, \hat{p}_{2,\phi,\psi}$ for $\hat{p}_1, \hat{p}_2$.

Many testing protocols are currently being used in HIV surveillance programs which aim to optimize efficiency and retain anonymity. There exists an ongoing debate about the ethics of unlinked anonymous testing (UAT) (Diaz et al., 2005; Krishnan and Jesani, 2009). In sentinel populations such as pregnant women at ANC clinics, UAT without informed consent is a commonly used protocol. Blood samples that are obtained for routine tests are also tested for HIV without any informed consent and are not linked back to the individual in any way. As treatment becomes more available, the ethics of such testing procedures become more questionable, and our suggested protocol requires obtaining informed consent from the individual. Voluntary UAT (or UAT with informed consent) is a much more widely accepted testing protocol and is currently used in DHS surveys. Informed consent is obtained before testing blood for HIV, but test results are not linked back to the individuals and, those who test cannot learn their disease status. Our testing protocol bypasses any of the ethical issues associated with UAT, as sampled individuals have three options: 1) test as an individual and learn their disease status, 2) test as an individual and do not learn their disease status, or 3) submit blood for pooled testing and

do not learn their disease status.

Preserving privacy of the pooled testers is a primary concern in our protocol. If a pool tests negative, we know the test results of individuals in the pool (negative) within the bounds of the sensitivity of the testing kit. Presumably, individuals are not as concerned with the confidentiality and identifiability of negative test results, and we are not concerned with this situation. If a pool tests positive, individual test results in the positive pool are non-identifiable mathematically for pools of size 2 or bigger. Of course, the issue of trust is important; those carrying out the survey need to convince those surveyed that their privacy requests be respected if we wish to lower q_3 as much as possible. Furthermore, ethical non-identifiability for positive pools may mandate larger pool sizes.

If a pool tests positive, the probability that an individual is positive (when $\phi = 1$) is $p/(1-(1-p)^k)$ by Bayes Theorem. For instance, when the population prevalence is 20%, the probability that an individual in a positive pool is HIV positive is 1 when $k = 1$ (individual testing), 0.56 when $k = 2$, 0.41 when $k = 3$, 0.34 when $k = 4$, 0.30 when $k = 5$, 0.27 when $k = 6$, and 0.25 when $k = 7$. Since the population prevalence is 20%, without testing at all, the probability a person is infected is 20%. As k increases, the probability that an individual tests positive given the pool tests positive approaches the population prevalence. Thus, as pool size and prevalence increase, we gain less additional information about the disease status of individuals in a pool when the pool tests positive.

However, using pool sizes that are too large decreases accuracy of the pooled testing estimator (Section 1.4.1). Hence, the key idea in this confidentiality protection problem is "to balance the need for confidentiality protection with legitimate needs of data users" (Cox and Zayatz, 1995). The United States' Federal Commission for Statistical Methodology lays out threshold rules for identifiability of survey responses for tabular data within U.S. Agencies; generally, at least 3-5 responses per cell are required for non-identifiability, but this minimum choice of responses per cell often varies with the sensitivity of the information and potential for disclosure (Federal Committee on Statistical Methodology, 1994). In order to use the pooled samples, pool size must be carefully selected by balancing the

precision of the pooled estimator with the ethical restraints imposed by nondisclosure of individual test information.

Lastly, in selecting survey design parameters, namely pool size and total sample size, an *a priori* estimate of q_2 is necessary. This proportion can be estimated by conducting a small pilot study in the population before the survey is conducted.

1.7 Statistical Properties of the Combined Estimator

In this Section, we provide descriptions of and proofs for the properties of the combined estimator, including the asymptotic unbiasedness, an analytic form of the variance estimate, the asymptotic distribution, and the finite sample bias.

1.7.1 Asymptotic unbiasedness of p_T

$$\begin{aligned} E(\hat{p_T}) &= E_Y(E(\tfrac{X_1}{n'} + \tfrac{Y_2}{n'}\hat{p}_2|Y)) \\ &= p_T \end{aligned}$$

1.7.2 Derivation of asymptotic variance of p_T

Recall $V_1 = p_1(1-p_1) = Y_1 Var(\hat{p}_1)$ and $V_2 = \frac{1}{k}(1-p_2)^2((1-p_2)^{-k}-1) = Y_2 Var(\hat{p}_2)$.

$$\begin{aligned} Var(\hat{p_T}) &= \underbrace{E(Var(\hat{p_T}|Y))}_{a} + \underbrace{Var(E(\hat{p_T}|Y))}_{b} \\ E(Var(\hat{p_T}|Y)) &= \tfrac{1}{n'}(q_1' V_1 + q_2' V_2) \\ Var(E(\hat{p_T}|Y)) &= \tfrac{1}{n'}(q_1' q_2' (p_1^2 - 2p_1 p_2 + p_2^2)) \\ Var(\hat{p_T}) &= \tfrac{1}{n'}(q_1' p_1(1-p_1) + q_2' \frac{1}{k}(1-p_2)^2((1-p_2)^{-k}-1) \\ &\quad + q_1' q_2' (p_1 - p_2)^2) \end{aligned}$$

1.7.3 Asymptotic Distribution of p_T

Define the new notation $V_1 = p_1(1-p_1)$ and $V_2 = (1/k)(1-p_2)^2((1-p_2)^{-k}-1)$. It follows that $n'q_1'Var(\hat{p}_1|Y_1) \xrightarrow{p} V_1$ and $n'q_2'Var(\hat{p}_2|Y_2) \xrightarrow{p} V_2$ and V_1 and V_2 are free of both Y_1 and Y_2.

Note that $\sqrt{n'q_1'}\hat{p}_1 \sim N(0, V_1)$ and $\sqrt{n'q_2'}\hat{p}_2 \sim N(0, V_2)$ are independent.

Further, $\sqrt{n'}(\hat{q}_1' - q_1', \hat{q}_2' - q_2')^T \sim N(0, q_1'(1-q_1')\mathbf{1})$, where $\mathbf{1}$ is a 2×2 matrix of 1s. Rewrite:

$$\sqrt{n'}(\hat{p}_T - p_T) = \sqrt{n'}(\hat{p}_1(\hat{q}_1' - q_1') + \hat{p}_2(\hat{q}_2' - q_2')) + \sqrt{q_1'}\sqrt{n'q_1'}(\hat{p}_1 - p_1) + \sqrt{q_2'}\sqrt{n'q_2'}(\hat{p}_2 - p_2))$$

Note that:

$$\sqrt{q_1'}\sqrt{n'q_1'}(\hat{p}_1 - p_1) + \sqrt{q_2'}\sqrt{n'q_2'}(\hat{p}_2 - p_2)) \sim N(0, q_1'V_1 + q_2'V_2).$$

and:

$$\sqrt{n'}(p_1(\hat{q}_1' - q_1') + p_2(\hat{q}_2' - q_2')) \sim N(0, q_1'(1-q_1')(p_1^2 - 2p_1p_2 + p_2^2)).$$

We know that $\hat{p}_1 \xrightarrow{p} p_1$ and $\hat{p}_2 \xrightarrow{p} p_2$. Applying Slutsky's rule, and the independence of X_i and Y_i, $i = \{1, 2\}$,

$$\sqrt{n'}(\hat{p}_T - p_T) \xrightarrow{L} N(0, V_{p_T})$$

where $V_{p_T} = q_1'p_1(1-p_1) + q_2'\frac{1}{k}(1-p_2)^2((1-p_2)^{-k}-1) + q_1'q_2'(p_1-p_2)^2$.

1.7.4 Derivation of finite sample bias in p_T

$$\begin{aligned} E(\hat{p_T}) &= E_Y(E(\tfrac{X_1}{n'} + \tfrac{Y_2}{n'}\hat{p}_2|Y)) \\ &= p_T + E\left(\tfrac{Y_2}{n'}\tfrac{k-1}{2(1-p_2)}var(\hat{p}_2)\right) + O\left((\tfrac{n'}{k})^{-\frac{3}{2}}\right) \\ &\approx p_T + \tfrac{k-1}{2n'k}(1-p_2)^{1-k}(1-(1-p_2)^k) + O((\tfrac{n'}{k})^{-\frac{3}{2}}) \end{aligned}$$

Revised Neonatal Tetanus Elimination Survey Protocol

Lauren Hund and Marcello Pagano

Department of Biostatistics
Harvard School of Public Health

2.1 Introduction

Since 1999, UNICEF, the United Nations, and WHO, along with other partners, have committed to achieving global elimination of neonatal tetanus (NT). Neonatal tetanus elimination is defined as less than 1 case of NT per 1000 live births in the highest risk district in a country. Detecting NT cases in the population is difficult using sentinel surveillance or household surveys. Sentinel surveillance of health facilities ignores any NT cases that were not taken to the facility for treatment. In household surveys, determining if an infant's illness was due to tetanus is difficult if the child was not taken to a health facility. Consequently, neonatal elimination surveys are conducted by using household surveys that monitor the NT mortality rate, defined as the number of deaths from neonatal tetanus per live births. The NT mortality rate is easier to monitor in practice due to the availability of the verbal autopsy method of diagnosis; further, the mortality rate among births with tetanus remains high in most countries conducting neonatal elimination surveys. An exception to this rule is China.

For a country to declare neonatal tetanus elimination, a lot quality assurance sampling survey is conducted to determine whether the nation achieved elimination. In this manuscript, we illustrate technical issues in the design of neonatal tetanus elimination surveys. We extend the work conducted in Stroh and Birmingham (2002), expanding the authors' ideas of using binary classification and double sampling designs to declare neonatal tetanus elimination in countries. Using the original survey design framework proposed in Stroh and Birmingham (2002), we detail statistical considerations pertaining to the survey design methodology.

In Section 2.2, we describe issues with identifying the target population, or highest risk district, for the survey. In Section 2.3, we introduce the lot quality assurance sampling (LQAS) methodology, and discuss extensions of this methodology that are used in NT surveys, including cluster sampling, finite population adjustments, and double sampling designs. In Section 2.4, we discuss the sensitivity of the survey instrument to detect

NT cases and the implications of monitoring a marker for NT incidence, namely the NT mortality rate. In Section 2.5, we describe the statistical calculations used to construct the survey design, and the metrics used to evaluate the properties of the design. Finally, in Section 2.6, we present a recommended survey design for the elimination of neonatal tetanus.

2.2 Selection of districts for the survey

NT elimination surveys occur at the district-level, where a district is defined as the third administrative level in a country (with nation as the first level). The first step in the design of an NT elimination survey is deciding which districts in a country are likely to have the highest NT incidence. These districts will be the target population for the survey(s). NT elimination is declared at the district level of aggregation, with the formal definition of NT elimination being "an NT incidence rate < 1 in 1000 live births in every district in a country." The implications of this definition are important to consider when designing an NT elimination survey.

NT rates less than 1 in 1000 live births must be achieved in every district. We only conduct elimination surveys in the highest-risk district(s), under the logic that if the rates of NT are less than 1 in 1000 in these districts, then the rates are also below this threshold in all the lower risk districts. This reliance on a prior ranking of NT risk within districts should be emphasized, since the validity of subsequently declaring a country as having achieved elimination depends on this ranking. Selection of the target population is non-trivial and a very important component of the survey design procedure.

If we can identify the worst performing districts with 100% accuracy, then only surveying the worst performing districts is an acceptable and accurate practice. However, if we have many districts with potentially high NT rates, then we need to survey all of these districts. We cannot randomly choose one or two of the districts to conduct surveys in, as we run the very real risk of failing to select a district that has not achieved elimination.

If we randomly select among high risks districts, we lose precision in our classification at the national level.

The smaller the district to be surveyed, the more precise we can be in our classification of elimination within that district (because, in small districts, we sample a large fraction of all the live births). On the other hand, if a district is too small, we may encounter operational problems. A case in point, when conducting an NT elimination survey, sometimes it is not logistically feasible to only survey the worst performing district, due to an insufficient number of live births in that district (e.g. such a survey might require sampling all of the live births in the district). In this situation, we can redefine the target population for an elimination survey by combining multiple high-risk districts into one survey. However, subsequent to this recombining, if we conduct a survey across districts, then we are changing the definition of elimination in this country, which should be clearly stated and approved by the assessment team before conducting the survey. The revised definition of elimination for the country is now "an average NT incidence rate < 1 in 1000 live births among the worst performing districts in a country."

In many situations, collapsing across multiple districts will be the most practical option. For instance, districts in Vietnam are frequently sub-divided, such that Vietnam had 34 districts in 1997, 125 by 2001, 424 by 2002 and 662 by 2011. It is impractical to conduct a survey at the district level in this situation; the number of live births per district is too small to construct a meaningful or logistically plausible sampling frame, and, further, the districts no longer represent meaningful subdivisions of the country.

It is important to not overlook the implications of pooling information across multiple districts in changing the definition of elimination. For instance, consider a situation in which we identify three high risk districts with a low number of live births. We decide to conduct one elimination survey, sampling from all three districts combined. Now, we define NT elimination in this country as "an average NT incidence rate < 1 in 1000 live births in the high risk districts." This definition is different from the standard definition of elimination that requires elimination in every district. Even if one of the three districts

has an NT incidence rate greater than 1 in 1000 live births, the country will usually be declared as having achieved elimination if the average incidence rate across the three high risk districts is less than 1 in 1000, since the average across these three districts can be less than 1 in 1000 without the rate being less than 1 in 1000 in each of the three districts.

2.3 Introduction to the LQA-CS survey methodology

The LQA-CS survey method is appropriate for selected populations in the final stage of MNT elimination when there is evidence suggesting that NT incidence has been reduced to less than 1 case/1000 live births and only occurs sporadically (not in clusters). Viewed as requiring a binary decision (has MNT elimination occurred, yes or no?), it is clear that no further requirement is made of the method to also provide an actual estimate of the MNT rate. In contrast, conventional surveys designed to estimate the NTMR with any degree of confidence require very large sample sizes - tens of thousands of live births - due to the extremely low incidence of NT in the final stages of MNT elimination (Dixon et al., 2005). Hence, the LQA-CS method is able to use relatively smaller sample sizes than the traditional estimation surveys (Valadez, 1991). Because of the smaller sample sizes required in general for the classification process (as opposed to the estimation process), the LQA-CS surveys are feasible and affordable in countries ready to demonstrate MNT elimination.

The LQA-CS survey assesses whether NT elimination in the target population has been achieved. Classification as having achieved or failed to achieve NT elimination is the goal, rather than estimation of the NTMR rate. NTMR rates can be estimated using LQA-CS data, but the estimates have large variances (resulting in very wide confidence intervals) and are susceptible to selection bias if the survey is stopped early. Therefore, calculating point estimates and confidence intervals for NTMR is not recommended; rather, the number of observed NT cases and the number of sampled live births should be reported.

In an LQA-CS survey, the number of NT deaths detected during the survey is compared to a pre-determined maximum acceptance number of NT deaths that defines whether the district "passes" (elimination achieved) or "fails" (elimination not achieved). The acceptance number is calculated to ensure that there is a high probability that a district with a high NT incidence rate during the 12 month interval covered by the survey does not "pass", and that districts with truly low NT rates do not "fail".

Lot quality assurance sampling (LQAS) survey designs in public health have been described extensively in the literature (e.g. Valadez (1991); Robertson and Valadez (2006)). We briefly describe the LQAS methodology, to aid in the interpretation of the final survey design.

2.3.1 Review of LQAS methodology

To declare elimination in a district, we need to decide whether the rate of neonatal tetanus mortality during the 12 month interval covered by the survey is sufficiently low. We denote the district-level NTMR as p. In the district, we sample n live births, and let X denote the number of cases of neonatal mortality caused by neonatal tetanus.

Assuming the population size/number of live births in a district is large ($> 50,000$), we can model X using a binomial distribution, specifically $X \sim Binomial(n, p)$. For some number d (the acceptance number), if $X > d$, we conclude that elimination has not occurred; if $X \leq d$, elimination has occurred. In choosing a sampling design for an LQAS survey, the goal is to select a sample size n and corresponding acceptance number d such that we run a small risk of misclassifying districts as having achieved or not achieved elimination. The lot quality assurance sampling (LQAS) survey design is determined by the following two equations, which control the error of the classification procedure:

$$P(X \leq d|n, p_u) \leq \alpha$$
$$P(X > d|n, p_l) \leq \beta$$

For a given choice of n and d, α is the probability that we classify a district as having

achieved elimination when the NTMR is greater than or equal to p_u; and β is the probability that we classify a district as not having achieved elimination, when the NTMR is less than or equal to p_l. To select an appropriate sample size n and decision rule d, we first need to decide what the relevant choices of p_l, p_u, α, and β are.

As an example, if we choose $\alpha = 0.1$ and $\beta = 0.1$, we then find a sample size n and acceptance number d such that we can make the following statement about our survey: "In an area with a true NTMR equal to 0.0021 (p_u) or more, if we repeat the MNTE elimination survey a very large number of times, we would incorrectly conclude that neonatal tetanus has been eliminated at most 10% (α) of the time. In an area with a true NTMR equal to 0.00035 (p_l) or less, if we repeat the MNTE survey a very large number of times, we would incorrectly conclude that elimination has not occurred at most 10% (β) of the time."

If a district has a true NTMR between p_l and p_u, we say that the NTMR lies in the "grey region." We do not restrict the classification errors within the grey region. Within this region, the risk of misclassification is higher than the smaller of α and β. To fully understand classification properties for districts with true NTMR in the grey region, we must examine the operating characteristic curve or the risk curve (see Section 2.5).

In the neonatal tetanus elimination surveys, we have selected p_l and p_u such that some districts with true NTMR rates in the grey region have not technically met the definition of elimination, but have achieved low enough NTMR rates that it is not a grave error to mistakenly declare elimination in these areas, if that mistake were to occur. Elimination surveys are only conducted when we have some confidence that elimination has been achieved, so ideally most districts will not have true NTMR rates that lie within the grey region. However, it is important to understand the inherent risk in the classification procedure.

In an LQAS survey, shortening the grey region results in more precise classifications. However, the length of the grey region is directly related to the sample size for the survey.

Table 2.1: Impact of the length of the grey region on the sample size. Single and double sampling plans are presented for large populations ($> 50,000$ live births per year) and a population with 5,000 live births per year.

	$> 50,000$ live births						5,000 live births					
	Single				Double		Single				Double	
p_u	d	n	d_1	n_1	d_2	n_2	d	n	d_1	n_1	d_2	n_2
3.0	2	2,540	0	1,430	2	1,310	1	1560	0	1,200	1	430
2.0	3	4,780	0	2,140	4	4,070	1	2270	0	1,740	1	650
1.5	5	8,840	0	2,860	6	8,000	1	2560	0	1,970	1	720
1.0	14	28,760	0	4,280	16	29,870	1	3400	0	2,640	1	920

When searching for a rare event in the population, required sample sizes are generally large, and we must balance precision and feasibility in our selection of p_l and p_u. Table 2.1 illustrates the impact of decreasing the grey region, lowering p_u when $p_l = 0.00035$, $\alpha = 0.1$ and $\beta = 0.1$ (as above). To convert the NTMR thresholds in Table 2.1 back to NT incidence rates, see the discussion about sensitivity and specificity in Section 2.4. Assuming sensitivity is 70% and specificity is 100%, these upper thresholds for NTMR correspond to NT incidence rates of 3, 2, 1.5, and 1 cases/1000 live births; and the lower NTMR threshold corresponds to an incidence rate of 0.5 cases/1000 live births

2.3.2 Finite population size effect

When the number of live births in a district is not sufficiently large ($< 50,000$ live births in the population), we model X using the hypergeometric distribution, $X \sim Hypergeometric(n, N, m)$, where n once again denotes the number of live births sampled. N is the total number of live births and $m = Np$ is the number of neonatal tetanus deaths in the district over the 12 month survey period for a given NTMR p. When N is large, the binomial and hypergeometric distributions are equivalent; the sample size and acceptance number for the survey will be identical regardless of which distribution is used for the calculations.

To design an LQAS survey, we calculate the parameter m using p_u and p_l. The con-

sequences of searching for a rare event in a finite population on the survey design are nontrivial. The NTMR p can only take on a finite number of values, since m is an integer by definition. Specifically, consider a population of 2,500 live births. The NTMR can only take on certain values in the population: $p = 0$ with 0 NT deaths; $p = 0.4/1000$ live births with 1 NT death; $p = 0.8/1000$ live births with 2 NT deaths; $p = 1.2/1000$ live births with 3 NT deaths; and so on.

When designing an LQAS survey, the grey region is usually no longer truly from p_l to p_u, but is wider, because p can only take on a finite number of values. For instance, in the example above with a population size of 2,000, if we select $p_l = 0.0005$ and $p_u = 0.002$, then the true grey region spans from 0.004 to 2, because p cannot take on the value of 0.0005.

The lengthening of the grey region impacts smaller population sizes more than the larger populations, where p can take on a wider range of values. It is important to discuss the appropriateness of the grey region when designing a survey with finite population sizes. For instance, if only 500 live births occur in a district, designing an elimination survey based on p_l and p_u is difficult. Elimination has only been achieved if 0 NT deaths occur in the district. The narrowest possible grey region is from 0 to 0.002, as p can only take on the values 0, 0.002, 0.004, etc. It is more intuitive and more appropriate to discuss absolute numbers of events, instead of focusing solely on rates, when dealing with very rare events in a finite population.

Given that NT is an endemic disease and cases can sporadically occur, the size of the target population should be sufficiently large to allow for the occurrence of acute, random cases without triggering an alarm signifying a chronic level. Therefore, the total number of eligible live births in a district should exceed 3,000 to conduct a meaningful NT elimination survey.

2.3.3 Cluster Surveys

Standard LQAS surveys usually select a simple random sample from the target population. Using simple random sampling requires an enumeration of the entire population in the district, sampling from this list, and then locating the sampled individuals. In MNTE surveys, it is impractical to implement simple random sampling within districts. Cluster sampling is logistically easier to carry out.

In NT elimination surveys, a cluster survey is conducted, but the data is analyzed by treating it as a simple random sample. Clusters for the LQA-CS survey are selected in the same manner as for a standard 30 x 7 cluster survey for immunization coverage (Lemeshow and Robinson, 1985). Note that the number of clusters and number of households to visit within each cluster in the LQA-CS survey are different from the 30 x 7 cluster survey.

As in the 30x7 surveys, probability proportional to size sampling is used for the selection of clusters. Specifically, the probability of a cluster being included in the survey is proportional to the number of live births in the survey.

Usually, cluster sampling increases the amount of variability in a survey, due to the fact that outcomes are more similar for individuals in the same cluster than for individuals in different clusters; so, to obtain a representative snapshot of the population, one needs to sample from many clusters. This within-cluster similarity is often quantified using the intracluster correlation coefficient (ICC or ρ); and the increase in variability in the survey estimators is measured by the design effect (DEFF), usually greater than one and defined to be the ratio of the variance of the survey using cluster sampling and the variance using simple random sampling.

To obtain the same level of precision with a cluster sample as one would obtain with a simple random sample, one needs to sample $n * DEFF$ individuals for the survey (often referred to as the effective sample size). When the number of clusters is large, and the population size within each cluster is large and approximately equal across clusters,

$DEFF \approx 1 + (m-1)\rho$, where m is the number of individuals sampled in each cluster.

When ρ is small relative to m, such that $DEFF \approx 1$, then we can treat the cluster sample like a simple random sample (and this is the current practice for neonatal tetanus surveys.) Historically, low design effects have been observed in surveys estimating NTMR (Rothenberg et al., 1985). Additionally, elimination surveys are only conducted when there is sufficient evidence that districts have low NT rates without any clustering of cases, adding credibility to the operative assumption. Lastly, from October 2000 to August 2011, the LQA-CS survey for the validation of MNT elimination has been implemented in 23 countries. One survey was conducted per country, except for India and Indonesia, who conducted 13 and 3 surveys, respectively. From the 41 survey reports available (where 4,571 clusters were visited), 42 neonatal deaths attributable to NT were reported. None of the clusters reported more than one neonatal death attributable to NT.

Therefore, we select households for inclusion in the survey using cluster sampling, and do not adjust for the impact of clustering in the survey (assume $DEFF = 1$), as we have this strong evidence that clustering effects are negligible in MNTE surveys; in other words, that $DEFF = 1$. If more than one NT death is found in a given cluster, efforts should be made to determine if the NT infections were related and due to a common cause or common risk factors. Specifically, if TT immunization rates vary substantially by cluster and unclean delivery and/or harmful cord care practices exist, clustering of NT cases is more likely. The cord care practice of one or several births attendants could also be the critical risk indicator for NT in absence of sufficient TT coverage in a cluster.

If the NT cases have a suspected common cause, such as the same birth attendant and no TT immunization, this important information should be presented and discussed when the final decision about NT elimination is made. On the other hand, if the clustering of NT cases is caused by lingering widespread use of hazardous delivery conditions or contaminated traditional substances on the cord, then the clustering could suggest that risky conditions and practices still exist in areas in which TT coverage is not sufficiently high.

2.3.4 Double sampling

A double sample procedure divides the total sample into two parts, and these parts are then surveyed sequentially - whether the second part is carried out is conditional on the results of the first part. This sampling procedure is analogous to interim monitoring in clinical trials. For additional sequential LQAS designs, see Myatt and Bennett (2008); Olives et al. (2009) for example.

Regardless of whether a single or double sampling plan is used, "failure to achieve elimination" can be declared at any point in the survey if the number of detected NT deaths surpasses the acceptance number, and the survey can be stopped early. If a large number of NT deaths are observed early in the survey, the survey should not be stopped until enough data (we require a representative sample of at least 250 mothers of eligible live births) has been collected to assess the remaining risk factors for NT (e.g. TT coverage; proportion of deliveries in a health facility and assisted by medically-trained attendants; and use of traditional substances on the umbilical stump).

It is important to keep in mind that, when the survey is stopped early, the collected data may not be representative of the entire district (because not all clusters have been visited). On many occasions, we may not want to stop the survey, even after the sample of 250 mothers was obtained. Specifically, if clusters are visited systematically (e.g. all urban clusters are visited first), then the collected data is susceptible to selection bias. Coverage estimates from the subsample obtained before the survey was stopped are no longer generalizable to the entire population. If clusters are visited on a random basis, the coverage estimates may be representative, even if the survey is stopped early. When representative coverage estimates of the additional indicators (e.g. vaccination, cord care, and clean delivery) are of interest, program managers must carefully consider whether the collected data is a representative sample of the district. If it is unclear whether the sample is representative, sampling should continue.

The double sampling plan has the advantage of allowing elimination to be declared

from the results of a preliminary first sample if the number of NT deaths detected is very low (e.g. 0). When the number of NT deaths in the first sample is not low enough to declare elimination (and the number of NT deaths in the first sample does not exceed the acceptance number), the second sample is necessary.

To construct a double sampling survey plan, we again specify thresholds p_l, p_u, α, and β. We also need to specify an additional parameter, α_1, which is the probability of declaring elimination after the preliminary sample, given p_u. This additional parameter does not affect the overall α-level of the survey design, but instead serves as a guide to select the sample size and decision rule for the preliminary sample. Based on these parameters, we can find the minimum sample sizes for the preliminary and secondary samples, n_1 and n_2, and the corresponding acceptance numbers d_1 and d_2, to meet our survey design specifications.

The proposed double and single sampling plans are designed using identical overall survey parameters p_l, p_u, α, and β. Therefore, to decide between a single and double sampling plan, we evaluate cost-effectiveness and feasibility, and are not concerned about the statistical precision of double versus single sampling (as they have the same precision). Thus the main reason that one would use a double sampling design is to reduce the amount of money/time spent conducting the survey.

Double sampling is only more cost-effective if we expect that the district has achieved elimination with some reasonable level of confidence. If the second sample is required, the total sample size required for a double sampling survey is always greater than the sample size for a single sampling survey. This result is due to the fact that we analyze the data twice during the survey period and have two different opportunities to declare elimination. In statistics, this issue is often referred to as "multiple comparisons", and we must adjust the classification errors to account for the fact that we look at the data twice. So, to obtain the desired classification errors α and β, we must sample more individuals in the double sampling plan to account for the inflated classification errors caused by looking at the data twice. As a general rule, we want to minimize the probability that we

will need the second part of the sampling.

Note that planning a double sampling survey also requires some extra effort when contrasted to a single sampling plan. Specifically, one must decide which clusters will be included in the first and in the second sample. Clusters should be divided between the samples such that the first sample is representative of the entire target population. Otherwise, inferences about the additional indicators (vaccination coverage, clean delivery, cord care, etc.) will not be representative of the surveyed population and will consequently be difficult to interpret. As an example, we cannot spatially partition the district to construct the first and second sample, though data collection would be much easier subsequent to such a partitioning. Additionally, one must analyze the data from the small sample, and decide whether the next sampling stage should occur. This interim analysis could be logistically challenging. Further, survey preparations are necessary for all clusters (for both the first and second sample) and may be considerable if a second part is required.

When choosing between a single versus double sampling plan, the deciding factor should be: "Is the cost/time savings that are potentially associated with double sampling worth the additional logistics that go into planning a double sampling survey and the potential extra cost of the second part?" So we need a measure of the odds that a second sample will be required. The odds of requiring a second sample decrease with the odds that the NT rate is well-below 1 in 1000 live births. If we expect that a second sample will be required in the double sampling plan (i.e. we are uncertain about whether or not elimination has been achieved), then we should choose a single sampling plan, to save both time and money. More commonly, it may simply be logistically infeasible to conduct a double sampling survey. For instance, a lack of communication equipment and/or long travel times between clusters would preclude the midpoint evaluation (to determine if the second sample is required).

In summary, the decision of whether to use a single or double sampling plan requires some prior information about the district-level NTMR and knowledge of the cost and

logistical differentials for single and double sampling plans.

2.4 Sensitivity, specificity and selection bias in mortality surveys

The definition of NT elimination is < 1 case of NT per 1000 live births. However, it is operationally easier to accurately monitor NT mortality, rather than detect actual cases of NT. We thus use NT mortality as a marker of what we ideally would like to measure, NT incidence. Consequently, we must consider the implications of measurement error induced by monitoring a proxy of our outcome of interest.

We can rephrase this issue in terms of the sensitivity and specificity of the survey instrument/protocol. In an NT survey, sensitivity is the probability that an NT case is detected, given that the NT case is included in the sample. Alternatively, we can state the sensitivity as the proportion of NT deaths in the sample that are detected by the survey instrument. An NT case can fail to be detected in two different ways: (1) the case is not fatal, or (2) the case is fatal, but NT is not deemed the cause of death.

NT cases are diagnosed using the verbal autopsy method (Anker et al., 1999). If we can assume that all deaths due to NT are diagnosed properly, then the sensitivity for our survey is equal to the mortality rate among cases of NT in the population. Once again, if the mortality rate is low, then we are using a very insensitive diagnosis for NT, and we need to adjust the survey parameters accordingly. Low sensitivity will result in possibly declaring that elimination has occurred, when it truly has not.

Selection bias and recall bias are also common issues in retrospective neonatal mortality surveys (Becker et al., 1993; Central Statistical Agency and ORC Macro, 2006; National Population Commission and ICF Macro, 2009; National Statistics Office and ICF Macro, 2009). When neonatal death rates observed in these surveys are lower than expected, we have evidence of non-sampling errors induced by selection and recall bias. Omission of live births and subsequent deaths for children who are not living at the time of the

interview is usually the most common source of non-sampling error in surveys of live births; children who die in early infancy are the most commonly omitted births. Additionally, households with live children are more likely to be suggested by the local guides, and houses with potential infant deaths are consequently bypassed. Some surveys have found that guides may also incorrectly displace child mortalities into the neonatal age group when under pressure to find NT deaths. Poor quality in the reporting of age at death could lead to under-reporting of infant deaths. Lastly, in some surveys, mothers with children were more likely to be at home at the time of the survey, as opposed to mothers without children, increasing the potential to miss additional neonatal deaths (Sokal et al., 1988). Selection bias could result in declaring that elimination has occurred when it has not.

Understanding the potential of non-sampling errors induced by selection and recall bias to impact the underestimation of NT incidence is important to obtaining accurate survey results. We can adjust the sensitivity of the survey instrument downward to account for this underestimation induced by these biases.

Specificity is the probability that a live birth included in the survey is correctly classified as not being an NT case. The specificity of the survey will be a function of the infant mortality rate and the specificity of the verbal autopsy method and should be close to 1 for NT surveys. If the verbal autopsy method for detecting NT deaths correctly confirms all non-NT deaths, then the specificity of the survey instrument is 1, and we do not misclassify any neonatal deaths as NT cases. Low specificity will result in possibly declaring that elimination has not occurred, when it truly has.

To adjust the survey design parameters p_l and p_u for the sensitivity and specificity of the survey instrument, we can exploit the relationship:

$$p = p^i x \text{sensitivity} + (1 - \text{specificity}) x (1 - p^i)$$

where p is the measured NT mortality rate using current survey protocol, and p^i is the true incidence rate of NT in the population.

The mortality rate among the births with NT sets an upper bound for the sensitivity. For example, if we assume that the mortality rate among the NT cases is 80%, then the highest possible sensitivity for the survey is 80%. In this case, we assume that NT mortality is high (80%), all cases of NT in the sample are detected, and selection and recall biases are not an issue. When NT mortality is lower, say 50%, and we expect that only 80% of NT deaths would ever be detected in the survey, then the sensitivity is $80\% * 50\% = 40\%$, and we need to adjust p_l and p_u downward by this 40%. Additionally, it is unreasonable to assume that recall and selection bias will not cause downward bias in NTMR estimates. Given that recall and selection biases impact most retrospective child mortality surveys, we should adjust the sensitivity further to reflect these biases.

It is clear that underestimating sensitivity is more conservative (i.e. harder to declare elimination) than overestimating sensitivity. Failing to adjust for sensitivity of the survey instrument will produce survey results that are difficult to interpret. It is much more likely that NT elimination could be incorrectly declared if the potentially low sensitivity of the survey instrument is ignored.

2.5 An explanation of probability calculations for operating characteristic curves

The LQA-CS method is considered the most practical for assessing whether MNT elimination has been achieved (Stroh and Birmingham, 2002). If districts at highest risk are surveyed and a pass decision is made, we conclude that other districts (at lower risk) have also achieved NT elimination (as discussed in Section 2.6).

The operating characteristic (OC) curve is defined as the probability of finding at most d (the acceptance number) NT deaths in the survey as a function of the true NT mortality rate in the district. To calculate the OC curves for a single sampling plan with sample size

n and acceptance number d, we use properties of the binomial distribution to calculate:

$$OC(p) = P(X \leq d|p) = \sum_{k=0}^{d} \binom{n}{k} p^k (1-p)^{(n-k)}$$

Recall that the OC curve is a function of p, the true NTMR rate in the district. The objective is to make the right tail of the OC curve as small as possible (minimize the probability of declaring elimination when p is large) and the left tail as large as possible (maximize the probability of declaring elimination when p is sufficiently small).

If the number of live births in the district is less than 50,000, because of the very low incidence of interest here, we recommend calculating the OC curve using the hypergeometric distribution (Section 2.3.2). The hypergeometric distribution accounts for the fact that the population size is finite (and is otherwise identical to the binomial distribution, which assumes an infinite population size). For populations with fewer than 50,000 live births, we calculate the OC curve using the formula:

$$OC(p) = P(X \leq d|N, m = Np) = \sum_{k=0}^{d} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

where $p = m/N$. Note that p can only take on a finite number of values when we use the hypergeometric distribution, since $m = \{0, 1, 2, , N\}$ is finite.

Calculations for the OC curve using a double sampling plan are slightly more complex. We design the surveys so that the probability of declaring elimination when $p > p_u$ is approximately the same for the single and double sampling plans. Equivalently, we say that the α-error of the single sampling plan is equal to the α-error of the double sampling plan. We also ensure that these plans have approximately equal β-errors.

To calculate an OC curve for a double sampling plan, we again need to calculate the probability that we declare elimination (pass) for a given rate of NT mortality in the population, but we need to consider the fact that we can declare elimination at two different time points. We calculate (1) the probability of passing at the first stage of sampling; and (2) the probability of passing at the second stage of sampling given that we did not pass at the first stage. To obtain the total probability of passing a district when using a

double sampling plan, we add these two probabilities (because the events are mutually exclusive).

$$\begin{aligned} OC(p) &= P(\text{pass}|p) \\ &= P(\text{pass at stage 1}|p) + P(\text{pass at stage 2 and not at stage 1}|p) \\ &= OC_1(p) + OC_2(p), \end{aligned}$$

where

$$\begin{aligned} OC_1(p) &= P(X_1 \leq d_1|p) = \sum_{k=0}^{d_1} \binom{n_1}{k} p^k (1-p)^{n_1-k} \\ OC_2(p) &= \sum_{k=d_1+1}^{d_2} P(X_1 = k|p) P(X_2 \leq d_2 - k|p) \\ &= \sum_{k=d_1+1}^{d_2} \binom{n_1}{k} p^k (1-p)^{n_1-k} \sum_{j=0}^{d_2-k} \binom{n_2}{j} p^j (1-p)^{n_2-j} \end{aligned}$$

Note that we first calculate the first stage sample size and acceptance number, n_1 and d_1, using thresholds p_l, p_u, α_1, and set $\beta_1 = 1$ (because we use the first sample to 'stop early' if we can declare elimination). Then, to finalize the second-stage sampling design, we calculate $OC(p)$ over a range of n_2 and d_2, fixing n_1 and d_1, searching for a sample size and acceptance rule with the pre-specified design properties. Then, using $OC(p)$, we examine whether the selected sample sizes and acceptance numbers meet the design specifications (governed by p_l, p_u, α, andβ).

Similar to the single sampling plan, we can use the hypergeometric distribution to calculate the OC curve for a double sampling plan when the number of live births in a district is less than 50,000 in the 12 month survey period. In this case, we would calculate

$OC_1(p)$ and $OC_2(p)$ using the hypergeometric as follows:

$$OC_1(p) = P(X_1 \leq d_1|p) = \sum_{k=0}^{d_1} \frac{\binom{m}{k}\binom{N-m}{n_1-k}}{\binom{N}{n_1}}$$

$$OC_2(p) = \sum_{k=d_1+1}^{min(n_1,d_2)} P(X_1 = k|p)P(X_2 \leq d_2 - k|p)$$

$$= \sum_{k=d_1+1}^{min(n_1,d_2)} \frac{\binom{m}{k}\binom{N-m}{n_1-k}}{\binom{N}{n_1}} \sum_{j=0}^{d_2-k} \frac{\binom{m-k}{j}\binom{N-n_1-(m-k)}{n_2-j}}{\binom{N-n_1}{n_2}}$$

2.5.1 Risk Curve

A closely related concept to the OC curve is the risk curve. The risk curve is a function that gives the risk of making a mistake in the classification. Its definition requires the same quantities as the OC curve, plus a cut-off point, p^*, to demarcate the acceptable NTMR from the unacceptable. Minimization of the risk curve is the desideratum of a good design. Plotting the risk curve clearly indicates the true NTMR at which we are most likely to "make an error" in declaring that elimination has or has not occurred (where elimination is defined as NT incidence < 1 case per 1000 live births. Adjusting p^* for the imperfect sensitivity and specificity of the survey, we define $p^* =$ sensitivity $* \, 1/1000 +$ $(1 -$ specificity$) * 1/1000 = 0.7/1000$ NT deaths per 1000 live births.

2.6 Choosing a sampling plan

To design an LQA-CS survey for NT elimination, we progress through the following steps.

1. Select p_l^i and p_u^i, the relevant upper and lower thresholds for an LQA-CS survey based on NT incidence. We select $p_l^i = 0.5$ cases/1000 live births and $p_u^i = 3$ cases/1000 live births.

2. Select error rates α and β. We select $\alpha = 0.1$ and $\beta = 0.1$. For the double sampling plans, we also choose $\alpha_1 = 0.05, \beta_1 = 1$.

   The choice of p_l^i, p_u^i and α and β is equivalent to stating: "In a district with a true NT rate equal to 0.003 (p_u^i) or more, if we repeat the MNTE elimination survey a number of times, we would incorrectly conclude that neonatal tetanus has been eliminated less than or equal to 10% (α) of the time. And in a district with a true NT rate equal to 0.0005 (p_1^i), if we repeat the MNTE survey a very large number of times, we would incorrectly conclude that elimination has not occurred 10% (β) of the time."

3. Adjust the thresholds p_l^i and p_u^i for the estimated sensitivity and specificity of the survey instrument (includes the mortality rate adjustment), to obtain new thresholds p_l and p_u. We assume that the sensitivity is 0.7 and specificity is 1, resulting in mortality thresholds $p_l = 0.35$ NT deaths/1000 live births and $p_u = 2.1$ NT deaths/1000 live births.

4. Calculate sample size based on α, β, p_l and p_u (and α_1 and β_1 for double sampling plans.). If the size of the target population is known and is less than 50,000 live births, we use the formulas based on the hypergeometric distribution for the calculations. Otherwise, we use the binomial distribution. Usually, the hypergeometric distribution will be more appropriate, as the target population of live births is usually substantially less than 50,000.

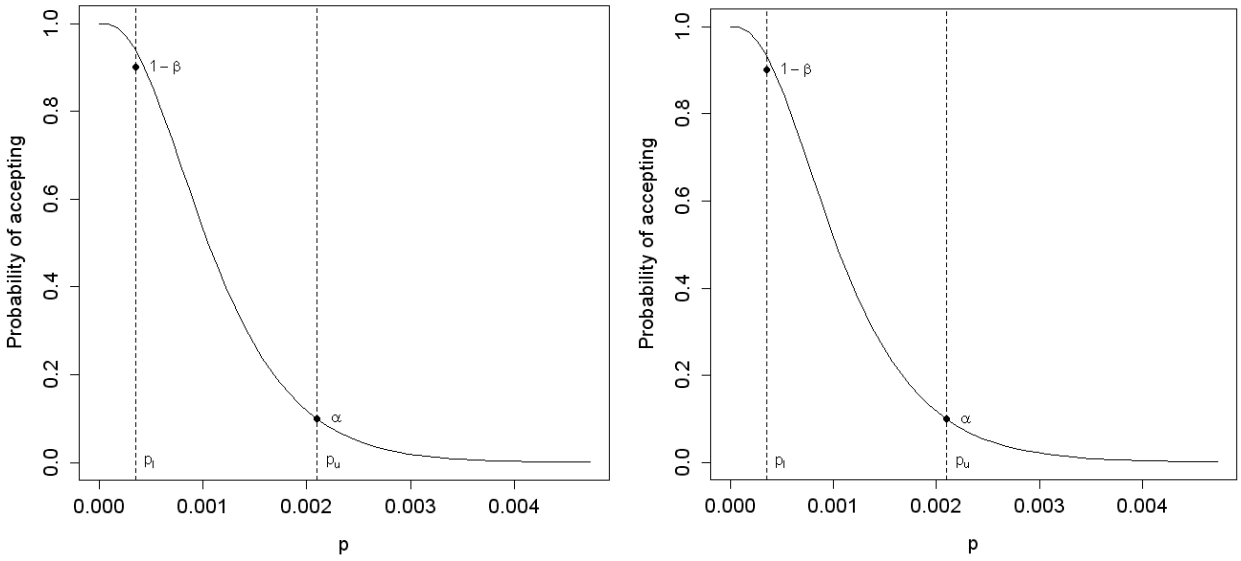
For a large target population ($> 50,000$ live births), we arrive at the following designs. When using a single sampling plan, we need to sample 2,540 live births, and declare elimination if we observe less than or equal to 2 cases of NT mortality.

With a double sampling plan, we should initially sample 1,430 live births. If we do not observe any cases of NT mortality, we declare elimination. If we observe greater than 2 cases, we declare elimination has not been achieved. If we observe exactly 1 or 2 cases, we sample an additional 1,310 live births. If we observe less than or equal to 2 cases among all $1,430 + 1,310 = 2,740$ live births, then we declare elimination. Otherwise, we

conclude NT elimination has not occurred.

The OC curves corresponding to these sampling designs are plotted in Figure 2.1. Note that the single and double sampling OC curves appear nearly identical, reflecting the fact that the single and double sampling plans were designed to have comparable statistical classification properties.

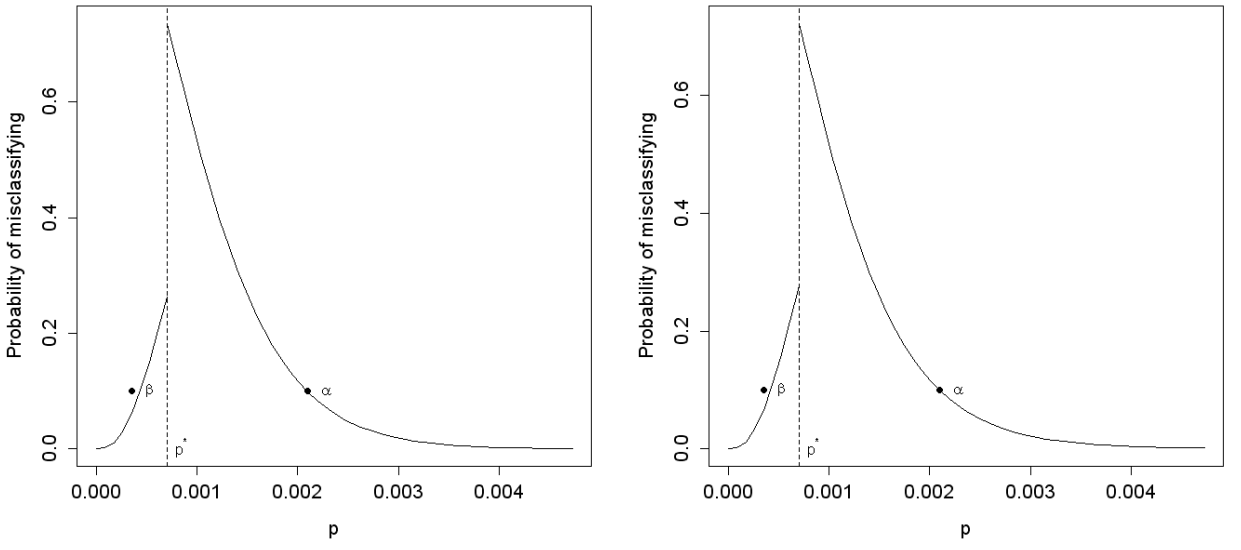
Figure 2.2 shows the risk curves corresponding to the OC curves in Figure 2.1 when $p^* = 0.7$ deaths/1000 live births. Using these figures, it is clear that the risk of misclassifying a district as having achieved elimination is high when the true NTMR in a district



(a) Single Sampling

(b) Double Sampling

Figure 2.1: OC curves for single and double sampling plans. Sample size and acceptance number calculated using $p_l = 0.00035$, $p_u = 0.0021$, $\alpha = 0.1$ and $\beta = 0.1$.



(a) Single Sampling

(b) Double Sampling

Figure 2.2: Risk curve for single and double sampling plans. Sample size and acceptance number calculated using $p_l = 0.00035$, $p_u = 0.0021$, $\alpha = 0.1$ and $\beta = 0.1$.

is between 0.7 and 2 NT deaths/1000 live births. We are willing to accept this risk, because an NTMR in this range is practically very close to achieving the formal definition of elimination and is consequently considered a major public health achievement for the country.

The risk of declaring that a country has not achieved elimination when it truly has remains relatively low ($< 30\%$). This property of the survey design is a consequence of choosing a value of p_l that is closer to p^* than p_u. If we select p_l and p_u such that they are equidistant from p^* (and choose $\alpha = \beta$), the risk of incorrectly declaring that a country has or has not achieved elimination should be close to 50% when the true NTMR p is very close to p^* (irrespective of whether it is higher or lower).

In Table 2.2, we list the probability of declaring that elimination has occurred, for various values of p (these are plotted in the OC curves as well, but are listed below for reference).

In Table 2.3, we present sample sizes and decision rules using the design parameters in Section 2.6, when the target population size is less than 50,000 live births.

Table 2.2: OC calculations for single and double sampling plans. Upper and lower thresholds are denoted with a *. Sample size and acceptance number calculated based on the parameters: p_l = 0.35 NT deaths/1000 live births and $p_u = 2.1$ NT deaths/1000 live births, $\alpha = 0.1$ and $\beta = 0.1$.

p (/1000)	Single	Double
0.0	1.000	1.000
0.1	0.998	0.998
0.2	0.985	0.984
0.35^*	0.939	0.934
0.5	0.864	0.855
0.7	0.737	0.723
1.0	0.534	0.519
2.0	0.118	0.117
2.1^*	0.099	0.099
3.0	0.018	0.022
4.0	0.002	0.004
5.0	0.0003	0.001

Table 2.3: Sample sizes for finite population sizes. $p_l = 0.35/1000$; $p_u = 2.1/1000$; $\alpha = 0.1$; $\beta = 0.1$; $\alpha_1 = 0.05$. In single sampling plan, sample n live births and denote number of NT deaths detected as X. Declare elimination if $X \leq d$. In the double sampling plan, sample n_i live births at stage i and denote number of NT deaths as X_i. Declare elimination when $X_1 \leq d_1$ and when $X_1 + X_2 \leq d_2$.

			Single Sampling				Double Sampling						
Pop	p_l	p_u	d	n	α	β	d_1	n_1	d_2	n_2	α_1	α_2	β
3,000	0.33	2.33	1	1,360	0.10	0.00	0	1,050	1	380	0.05	0.10	0.00
4,000	0.25	2.25	1	1,480	0.10	0.00	0	1,140	1	410	0.05	0.10	0.00
5,000	0.20	2.20	1	1,560	0.10	0.00	0	1,200	1	430	0.05	0.10	0.00
6,000	0.33	2.17	1	1,610	0.10	0.07	0	1,240	1	450	0.05	0.10	0.07
7,000	0.29	2.14	1	1,650	0.10	0.06	0	1,270	1	470	0.05	0.10	0.06
8,000	0.25	2.12	1	1,690	0.10	0.04	0	1,300	1	470	0.05	0.10	0.04
9,000	0.33	2.11	1	1,710	0.10	0.10	0	1,320	1	480	0.05	0.10	0.10
10,000	0.30	2.10	1	1,730	0.10	0.08	0	1,330	1	490	0.05	0.10	0.08
15,000	0.33	2.13	2	2,370	0.10	0.03	0	1,340	2	1,220	0.05	0.10	0.03
20,000	0.35	2.10	2	2,440	0.10	0.04	0	1,380	2	1,250	0.05	0.10	0.05
25,000	0.32	2.12	2	2,440	0.10	0.04	0	1,380	2	1,240	0.05	0.10	0.04
30,000	0.33	2.10	2	2,470	0.10	0.04	0	1,400	2	1,260	0.05	0.10	0.05
40,000	0.35	2.10	2	2,490	0.10	0.05	0	1,400	2	1,290	0.05	0.10	0.06
50,000	0.34	2.10	2	2,500	0.10	0.05	0	1,410	2	1,280	0.05	0.10	0.05

LQAS survey designs for monitoring the prevalence of malnutrition

Lauren Hund[1], Megan Deitchler[2], and Marcello Pagano[1]

[1] Department of Biostatistics
Harvard School of Public Health

[2] Food and Nutrition Technical Assistance Project
Academy for Educational Development

3.1 Introduction

Lot Quality Assurance Sampling (LQAS), also referred to as sampling for attributes or acceptance sampling, has a long history of applications in industrial quality control (Dodge and Romig, 1929). In the past 20 years, LQAS applications have become increasingly popular in global health care surveys (Robertson and Valadez, 2006).

Recently, LQAS cluster survey designs were introduced to classify prevalence of acute malnutrition as acceptable or high in emergency settings (Deitchler et al., 2007, 2008). LQAS malnutrition surveys were criticized for reporting too many false positives (classifying areas of acceptable malnutrition status as unacceptable) (Bilukha, 2008; Bilukha and Blanton, 2008). The poor classification properties of LQAS surveys were claimed many years ago by Sandiford (1993) in the context of vaccination coverage. To aid in the interpretation of malnutrition surveys, Bilukha and Blanton (2008) suggest reporting probability of high malnutrition in an area within the study results. In response, Olives and Pagano (2010) illustrate the difficulties in reporting false positive rates and illustrate how a Bayesian methods must be used to achieve this objective.

Additionally, existing LQAS malnutrition survey designs require sampling a large number of clusters to minimize the impact of within-cluster correlation. The cost-effectiveness and feasibility of survey designs that require visiting over 60 different clusters is questionable.

In this paper, we review LQAS survey designs for monitoring global acute malnutrition and propose extensions to the existing designs to address limitations in LQAS malnutrition surveys. In Section 3.2, we review LQAS surveys for monitoring the prevalence of malnutrition in children. In Section 3.3, we propose a simple adjustment to sample size calculations for LQAS surveys to incorporate within-cluster correlation. In Section 3.4, we draw from the historical quality control literature to introduce a framework for incorporating LQAS into longitudinal surveillance systems for acute malnutrition. This framework provides principled guidelines for designing an LQAS-type classification pro-

cedure to detect changes in malnutrition prevalence over time in a region.

3.2 Review of LQAS surveys for malnutrition

Malnutrition is frequently quantified using the binary indicator global acute malnutrition (GAM), usually defined as a weight-to-height Z-score (WHZ) < -2 and/or bipedal edema; alternatively, GAM is defined as a middle-upper arm circumference (MUAC) < 125mm and/or presence of an edema. The World Health Organization classifies malnutrition prevalence as critical if the prevalence of GAM in a population is $\geq 15\%$ (World Health Organization, 2000). Severity of malnutrition in an area is often assessed using surveys of GAM prevalence in children age 6-59 months (Deitchler et al., 2007). It is important to accurately classify the prevalence of malnutrition as high and to detect sudden rises in malnutrition prevalence using cost-effective surveys in order to inform when aid should be sent to a region and how resources should be allocated to reduce malnutrition.

The prevalence of acute malnutrition has traditionally been assessed using 30x30 cluster sampling surveys (30 clusters of 30 children) (Binkin et al., 1992), though there is currently no general consensus as to the optimal survey design to assess the prevalence of acute malnutrition (Spiegel, 2007). Deitchler et al. (2007) propose using LQAS surveys with cluster sampling to assess the prevalence of global acute malnutrition (GAM) based on WHZ scores and field test 33x6 and 63x7 cluster survey designs using pre-specified classification thresholds. LQAS surveys are typically less costly than traditional 30x30 cluster survey designs for estimating the prevalence of malnutrition, due to the smaller sample sizes required (Deitchler et al., 2008).

We review the LQAS malnutrition surveys presented in Deitchler et al. (2007). In a study region, community health workers collect measurements on n children, and find that X out of the n children have GAM. The number of children with GAM is then modeled using the binomial distribution, $X \sim Binomial(n, p)$, where p is the true proportion in the surveyed area. For some number d, if $X > d$, malnutrition prevalence is classified

as high; if $X \leq d$, then the malnutrition prevalence is classified as acceptable.

In choosing a sampling design for an LQAS survey, the goal is to select a sample size n and decision rule d such that we run a small risk of misclassifying districts as requiring intervention or not. The LQAS survey design is determined by the following two equations, which control the risk profile of the classification procedure:

$$\begin{aligned} P(X > d | p \leq p_l) &\leq \beta \\ P(X \leq d | p \geq p_u) &\leq \alpha \end{aligned}$$

We want to minimize the risk of classifying prevalence as high when the true prevalence of GAM is "low," and minimize the risk of classifying prevalence as acceptable when the true prevalence is "high." The meanings of "low" and "high" are determined by the choice of p_l and p_u, the lower and upper thresholds, chosen based on contextual knowledge. To design a survey, we specify α, the probability of classifying as acceptable when the true GAM prevalence is greater than p_u; and β, the probability of classifying prevalence as high when the true GAM prevalence is less than p_l.

Policy-makers decided to select $\alpha = 0.1, \beta = 0.2$ as acceptable risks, and selected 3 couplets for p_l and p_u: (1) 5-10%, (2) 10-15%, and (3) 15-20%. Based on these design features, 33×6 ($n = 198$) and 67×3 ($n = 201$) cluster sampling designs were chosen as guide designs for monitoring malnutrition prevalence, with respective decision rules 13, 23, and 33 (Deitchler et al., 2007). (Note that these designs have classification risks that are near the $\alpha = 0.1$ and $\beta = 0.2$, but do not necessarily meet these cut-offs).

Due to the infeasibility of implementing simple random sampling in emergency settings, children are sampled within villages (cluster sampling). Olives et al. (2009) demonstrate via simulation that these survey designs result in negligible clustering effects, because the cluster sample sizes are small (size 3 or 6) and the number of clusters sampled is large. For traditional 30×30 surveys, the effects of clustering vary by region (Katz, 1995).

Table 3.1: Average GAM prevalence, by WHZ (W) and MUAC (M). Number of surveys (Surv.) at each location; and number of kids, households (HH), and clusters sampled (Clus.); and average age of the sampled children in months are shown. Locations (Loc.) are Garissa (G) - pastoral (P), riverine (R), and urban (U); Mandera (M) - pastoral (P), riverine (R), and urban(U); Mathere Slum (MS), Sudan (S) - urban (U) and riverine (R).

Loc.	Surv.	Kids	HH	Clus.	GAM-W	GAM-M	Age
GP	3	224.7	163.0	33.0	15.5	3.8	30.3
GR	3	225.7	168.0	33.0	16.8	2.7	31.2
GU	3	231.3	170.0	33.0	14.6	5.4	31.7
MP	3	226.7	181.0	33.0	27.9	13.4	27.4
MR	3	221.7	151.7	33.0	36.6	12.3	30.6
MU	3	234.7	161.3	33.0	10.3	8.4	29.5
MS	3	230.7	139.0	33.0	21.4	6.1	30.8
SU	4	317.5	177.0	32.5	22.3	10.5	30.2
SR	2	222.5	137.0	33.0	21.2	7.7	30.9

Table 3.2: Number of surveys with high prevalence classification, out of 28 total surveys.

$p_l - p_u$ Couplet	High Gam
5-10%	26
10-15%	25
15-20%	19

3.2.1 LQAS surveys for monitoring malnutrition in Kenya and the Sudan

LQAS 33×6 surveys were conducted at three sites in South Sudan and seven sites in Kenya at six month intervals during 2008 and 2009. Most sites have conducted three rounds of surveys. Table 3.1 contains summary statistics for the surveys conducted at each location. Trends in GAM prevalence by location across time are shown in Figure 3.1. Prevalence of malnutrition is high in each location, peaking in the Mandera region. The estimates of GAM prevalence differ substantially when MUAC, rather than WHZ score, is used to construct the GAM indicator. Following standard protocol for LQAS malnutrition surveys, we use WHZ score to construct the GAM indicator in our analyses.

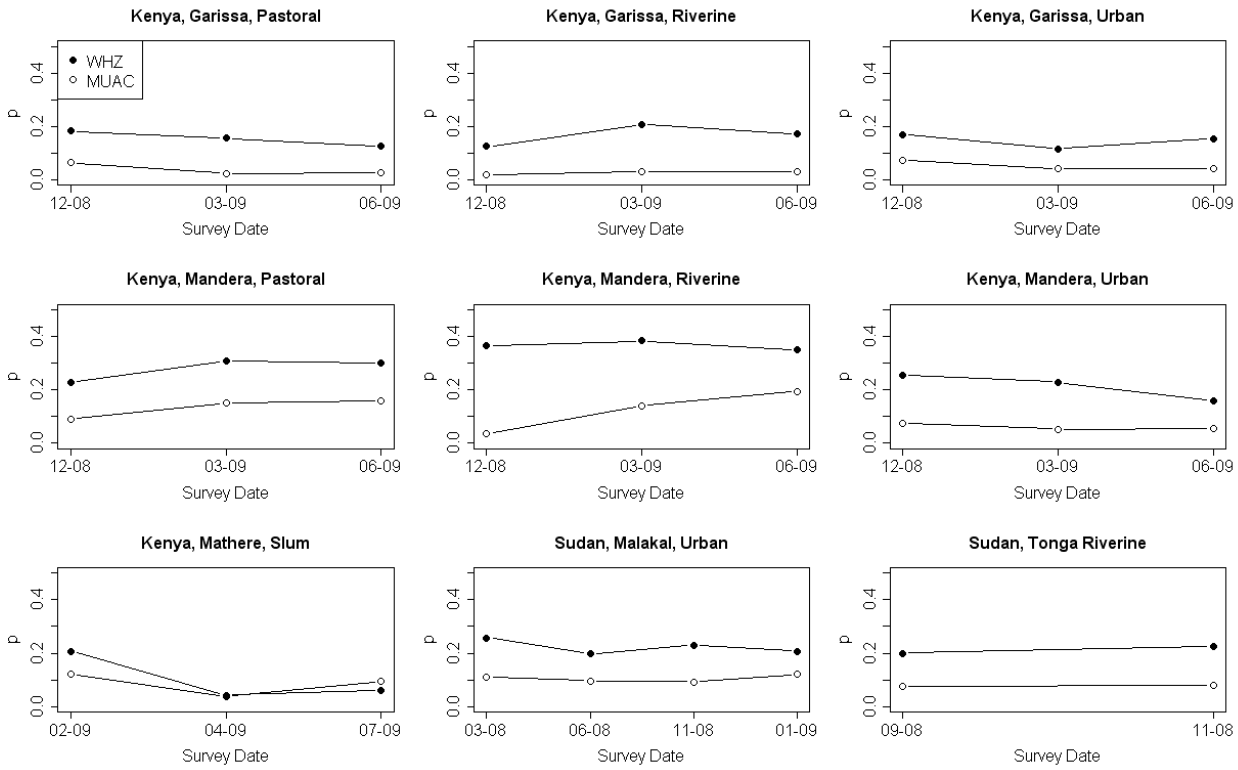


Figure 3.1: Prevalence of GAM by location

Using the pre-specified LQAS malnutrition survey designs, we analyze the results of 28 different 33×6 LQAS surveys according to the survey protocol. Because all of the surveys included more than 198 children (most had ~ 220 children in the survey), for sake of this discussion, we randomly deleted observations from the surveys to produce samples of 198 and used the standard LQAS decision rules. We repeated this procedure 100 times, and determined on average, in how many surveys we classified the prevalence of GAM as high, based on the 3 couplets above. Results of the data analysis are presented in Table 3.2. The prevalence of GAM has seasonal and regional variations, but is chronically high in most regions.

3.3 Incorporating clustering

Current LQAS cluster sampling designs for monitoring GAM assume that classification risks calculated under simple random sampling (SRS) are preserved in the cluster sam-

pling design. Assuming individuals within clusters are similar, cluster sampling will inflate classification risks, with the amount of inflation depending on the sample size per cluster and on the intracluster correlation (Lohr, 1999).

Define K as the number of clusters in the population; k as the number of clusters sampled; M as the population size within each cluster (assumed equal across clusters); and m as the sample size per cluster. The intracluster correlation quantifies the magnitude of the between cluster variability in prevalence relative to the within cluster variability, and can be defined as

$$\rho = 1 - \frac{M}{M-1}\frac{SSW}{SSTO},$$

where M is the within-cluster population size; SSW is the within-cluster sum of squares; and $SSTO$ is the total sum of squares (Lohr, 1999).

Recently, several methods have been proposed for preserving classification risk when clustering is present (Pezzoli et al., 2009; Greenland et al., 2011; Hedt-Gauthier et al., 2012). Results from previous surveys suggest that ρ may be low enough in the malnutrition setting that current LQAS designs are valid (Deitchler et al., 2008), while other studies inconclusively suggest clustering of malnutrition status exists at the household- or village-level (Fenn et al., 2004; Katz, 1995).

Simulation studies verified that current LQAS designs in the malnutrition setting preserve classification risks (Olives et al., 2009). However, making this strong parametric assumption and relying on a 'low enough' ρ could lead to problems when ρ is high. Additionally, the cost effectiveness of the current designs are questionable, because they requiring visiting a large number of clusters, which may be infeasible in practice (Binkin et al., 1992).

We propose a simple design procedure to incorporate clustering. Estimates from cluster sampling survey designs have higher variances than those from simple random samples. Following Rao and Scott (1992), we exploit the relationship between the intraclass correlation (ρ) and the design effect to adjust the effective sample sizes used. The effec-

tive sample size is the sample size required if we were to take a simple random sample from the population to achieve the same variance as in the cluster survey design (Rao and Scott, 1992).

We assume that (1) the number of individuals sampled per cluster m is constant, (2) the population size within each cluster M is large and is equal across clusters, and (3) the number of clusters in the population K is large. (These assumptions are identical to those in Hedt-Gauthier et al. (2012) and Pezzoli et al. (2009)). Then, the design effect is (Kerry and Martin Bland, 2001):

$$DEFF = 1 + (m - 1)\rho.$$

When the number of clusters in the population K is small (i.e. we sample a significant fraction of the clusters in the population), the design effect is approximately:

$$DEFF = 1 + (fm - 1)\rho$$

where f is the first stage finite population correction, $(1 - k/K)$ (see Section 3.8.1 for derivation).

To design an LQAS survey with cluster sampling, we iterate through choices of m and k until we find a decision rule that meets the classification risks α and β for a given choice of p_l and p_u. The algorithm proceeds as follows:

1. For a given m and k, calculate the effective sample size , $n^* = n/DEFF$, where $n = mk$. Round n^* to the nearest integer.

2. Search for a decision rule d^* using the standard binomial LQAS model with sample size n^*, specifying α, β, p_l, and p_u.

3. To obtain the final decision rule, calculate d^*DEFF and round this quantity to the nearest integer, to obtain the decision rule d.

4. The final sample size is n, consisting of k clusters of size m, and the decision rule is d.

For a given k, we are not guaranteed that sample size and decision rule exist that meet the risk thresholds, α and β. To find the minimum number of clusters that must be sampled, we consider the properties of the effective sample size as m (the within-cluster sample size) increases. Using this formulation, we see that as m gets large, the effective sample size converges to

$$n^*_{max} = \frac{k}{(1 - \frac{k}{K})\rho}$$

(see Section 3.8.2 for derivation). We must sample enough clusters such that a decision rule exists for n^*_{max} that meets the design specifications under simple random sampling using the binomial model.

This method for sample size calculation extends the existing methods (e.g. Pezzoli et al. (2009); Hedt-Gauthier et al. (2012)) by allowing for a finite number of clusters in the population. The method is conceptually closer to Hedt-Gauthier et al. (2012), relying on specification of the intraclass correlation ρ, rather than the standard deviation of p, $sd(p)$ (Pezzoli et al., 2009). All of these methods will produce similar results when p_l and p_u are bounded away from 0 or 1, but differ when p_l and p_u are close to 0 and 1. Fixing ρ as a constant instead of $sd(p)$ guarantees that the support of p is always between 0 and 1.

The major limitation to our clustering adjustment is the potential for rounding errors to inflate the classification risks, α and β. Because we have rounded in steps (1) and (3), our procedure is inexact and classification risks will generally be close to, but not exactly equal to, α and β. Rounding generally slightly increases one of α or β, but not both (see Figure 3.3). Additionally, as in Hedt-Gauthier et al. (2012), an estimate of ρ is needed to design the survey. In an ongoing longitudinal surveillance program, we can update estimates of ρ over time.

We no longer need to stay within the confines of the 67×3 or 33×6 designs to ensure that clustering does not inflate our classification risks. Insofar as we can obtain a reasonable estimate of ρ, we can design surveys that meet the classification risks for various combinations of m and k. For instance, suppose we know that approximately 20 children can be sampled in a cluster per day. Then, we could fix $m = 20$, and use

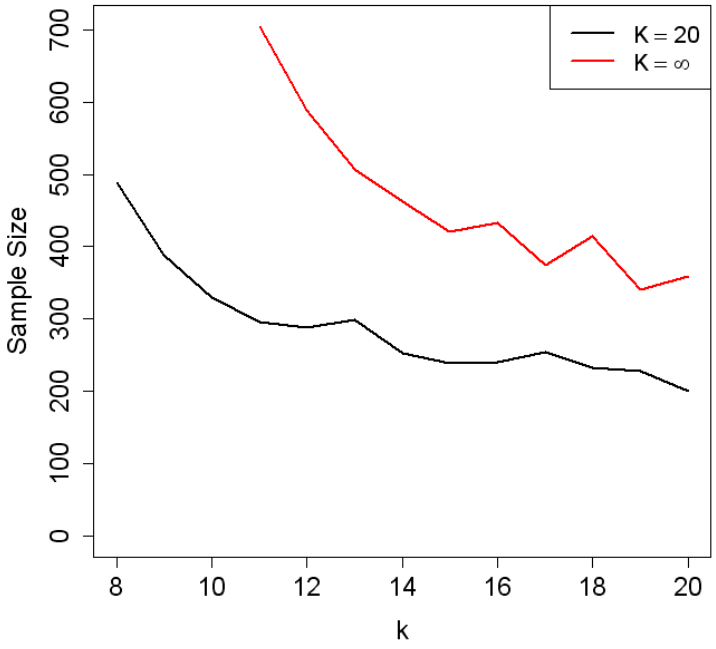
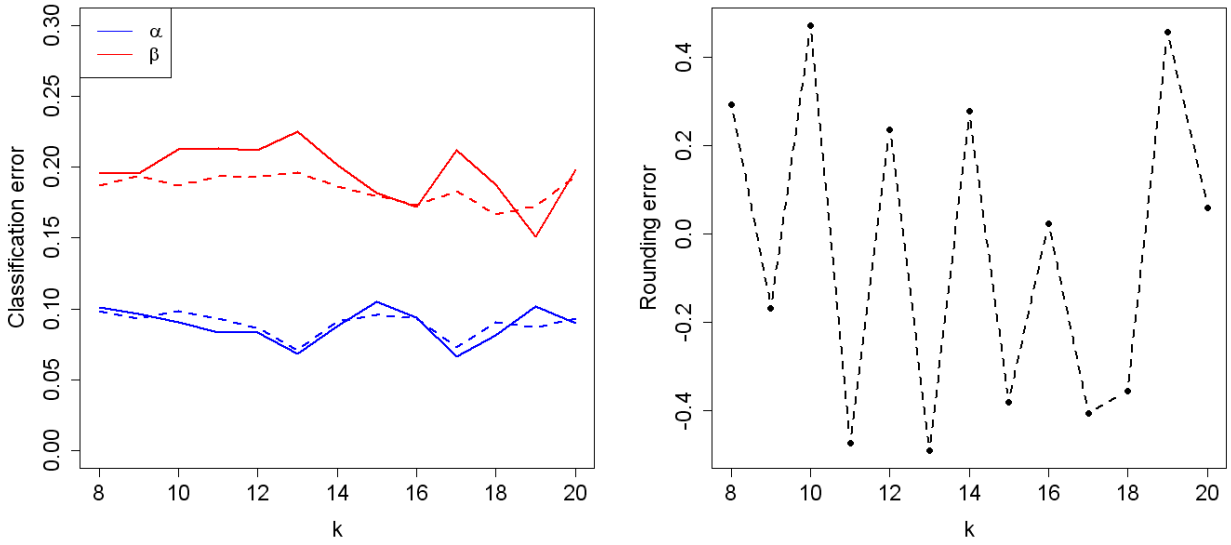


Figure 3.2: Sample sizes for LQAS survey designs when $p_l = .1, p_u = .15, \alpha = .1, \beta = .2$, comparing finite($K = 20$) versus infinite number of clusters, when $\rho = 0.05$.

the design procedure above to determine the number of total clusters that we need to visit. Alternatively, we could compare the expected cost of different choices of m and k to decide on a final design (Hedt-Gauthier et al., 2012).

To illustrate the performance of the design effect correction for clustering, we evaluate the properties of the LQAS survey design with $p_l = 0.10$, $p_u = 0.15$, $\alpha = 0.1$, and $\beta = 0.2$ in simulation. To construct our two simulated populations, we generate $K = 20$ cluster level prevalences from a Beta distribution, with correlation coefficient $\rho = 0.05$ and mean prevalences equal to p_l and p_u. We then scale the 20 prevalence estimates so that their means are exactly equal to p_l or p_u, and their intracluster correlations are exactly equal. Therefore, the prevalence estimates within clusters no longer follow a Beta distribution, but have the correct means and intracluster correlations.

Results of the simulation are shown in Figures 3.2 and 3.3. Assuming the number



(a) Classification risks

(b) Rounding error

Figure 3.3: Classification risks and rounding error for finite cluster LQAS survey designs when $p_l = 0.10, p_u = 0.15, \alpha = 0.1, \beta = 0.2$, $K = 20$, and $\rho = 0.05$. Panel (a): empirical (dotted line) and calculated (solid line) classification risks. Panel (b): rounding error for the decision rule ($d^* DEFF - d$).

of clusters in the population is infinite will result in a much larger sample size than necessary, when the true number of clusters in the population is small (Figure 3.2). In the simulated data, the estimated classification risks are close to the empirically calculated risks, and differences between the estimated and empirical risks are driven by rounding error (Figure 3.3). Rounding error results in an increase in one of α or β; and a decrease in the other.

3.4 Designing surveillance tools to detect changes over time

How to design and interpret the results of an LQAS survey for malnutrition depends on the goals of the survey. We consider two different motivations for conducting an LQAS survey.

First, consider an LQAS design to determine if regions are meeting pre-specified guidelines for intervention (Setting 1). For instance, the WHO recommends setting up therapeutic feeding centers in populations with GAM prevalence in children 6-59 months greater than 10% (World Health Organization, 1999). Implementing LQAS in this framework is straightforward, choosing p_l and p_u such that feeding centers are set up in populations with prevalence p_l or lower with probability less than or equal to α, and feeding centers are not set up in populations with prevalence p_u or higher with probability less than or equal to β. We can use the LQAS protocol described in Sections 3.2 and 3.3 to design these surveys. For instance, we might choose $p_l = 5\%$ and $p_u = 10\%$, acknowledging the fact that the risk of intervening in areas with prevalences between 5-10% is greater than α and increases as the prevalence approaches the upper threshold 10%.

Alternatively, for programs with longitudinal surveillance of malnutrition, the goal of the survey might be to detect changes in the malnutrition prevalence signalling a malnutrition crisis (Setting 2). If we observe a substantial spike in the malnutrition prevalence in a population, we need to quickly intervene. In this section, we discuss designing LQAS surveys for detecting spikes in the malnutrition rates (Setting 2). For now, we assume that the effects of clustering are negligible (*i.e.* we collect a simple random sample at each time point). In Section 3.4.2, we address how to adjust for using a cluster sampling design at each time point.

In the manufacturing industry, the distinction between Setting 1 and Setting 2 is analogous to the difference between quality control and process control (Sower et al., 1993). Quality control is concerned with balancing the number of defective goods sold to the consumer with the cost of repairing defective goods for the producer. Detecting rises in malnutrition prevalence is inherently tied to statistical process control (Colosimo and Del Castillo, 2006). In process control, an indicator (e.g. malnutrition prevalence) is tracked over time, and an alarm is sounded when the indicator goes 'out of control' (when a spike in malnutrition prevalence occurs). Knowledge of when a process is 'in control' (baseline acceptable malnutrition prevalence) is necessary for understanding when the

process is 'out of control.'

This baseline acceptable rate of malnutrition varies across populations, due to differences in body-types or different definitions of GAM. For instance, pastoralist populations tend to be tall and thin. Relatively healthy children in pastoralist population are more likely to be classified as malnourished than those in other populations, when GAM is defined using WHZ scores (Myatt et al., 2009).

When this baseline acceptable rate is unknown, we may be able to specify a range of rates that are acceptable for a given population, *e.g.* between 0-5%. Equivalently, we specify a baseline distribution of acceptable malnutrition rates, denoted $f_0(\cdot)$. We then compare this 'in-control' baseline distribution to the data that we collect to determine if we have observed a spike in malnutrition rates.

Our survey design procedure for detecting spikes in malnutrition in a population where the baseline rate is unknown is motivated by Yousry et al. (1991), who suggest using empirical Bayes process control theory to monitor the defect rate using binary indicators when the baseline in-control rate is unknown. We apply this general approach to aid in the design of LQAS surveys for monitoring GAM.

While our survey design uses empirical Bayes principles to estimate the baseline distribution, sample size and decision rule calculations are based on classical acceptance sampling theory. The distinction between Bayesian acceptance sampling (Olives and Pagano, 2010) and classical acceptance sampling is described in Section 3.7.

When designing surveys to detect sudden rises in malnutrition, we consider four different scenarios:

1. the baseline rate of malnutrition, p_0, is known,
2. the baseline distribution of malnutrition $f_0(\cdot)$ is known,
3. we have some historical information about the baseline distribution of malnutrition and have data from a baseline survey,

4. we have some historical information about the baseline distribution of malnutrition and have data from k surveys.

Following the initiation of a longitudinal surveillance program, we anticipate that Scenario 3 will hold, and we will not know much about the population of interest. Over time, we gather more information about the population of interest (Scenario 4), until we have a stable estimate of the baseline rate of malnutrition, with uncertainty bounds (Scenario 2). As we gather more information, we will may be able to use a known baseline rate of malnutrition p_0 to detect spikes in prevalence (Scenario 1).

3.4.1 Detecting deviations from a baseline distribution

Scenario 1

Scenario 1 is easy to accommodate in practice. We conduct a survey at time t and observe X_t out of N_t malnourished children. We can then use standard LQAS protocol, selecting a lower threshold $p_l = p_0$ and an upper threshold $p_u = p_0 + \Delta_U$, where Δ_U is a meaningful deviation in prevalence from the baseline. In this setting, the baseline distribution $f_0(\cdot)$ is a point mass at p_0.

Scenario 2

To design a survey when the baseline distribution of malnutrition $f_0(\cdot)$ is known (Scenario 2), we model X_t using a betabinomial distribution. That is, $f_0(\cdot)$ is a Beta distribution, and assess whether the observed data X_t is consistent with $f_0(\cdot)$ shifted by Δ_L or by Δ_U. (Figure 3.4).

Consider the following example. In a population, the prevalence of malnutrition has historically varied between 3% and 6% due to random fluctuations; when the prevalence is within this range, the population is considered relatively well-off. We assume the baseline distribution follows a Beta distribution with mean equal to $(3 + 6)/2 = 4.5\%$, with

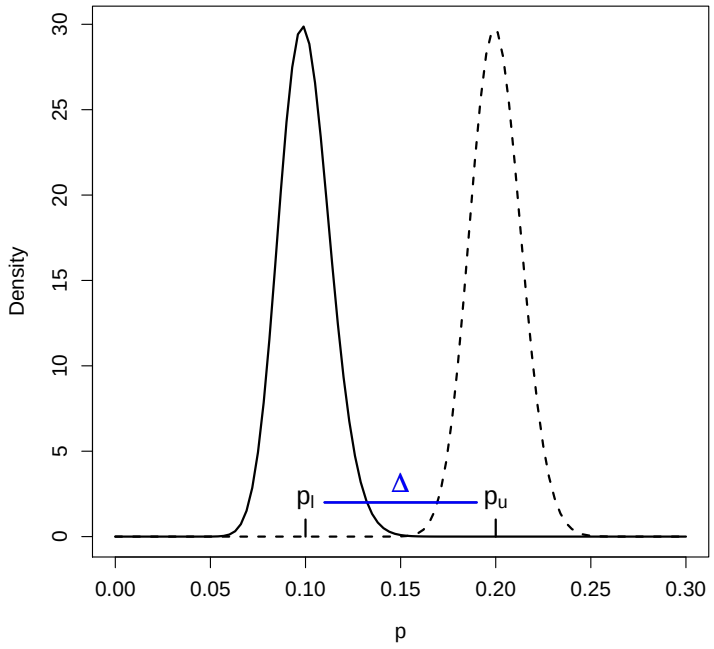


Figure 3.4: Searching for an increase Δ in p_{t-1} at the current time point, when p_{t-1} is measured with error.

95% of the density between 3 and 6%. Using these specifications, we estimate $a = 34.6$ and $b = 733.7$.

When prevalence is low at time t, $p_t \sim f_0(\cdot), X_t \sim Betabinomial(n_t, a, b)$, and we can calculate $P(X_t > d|f_0(\cdot))$ for a given n_t and d. Next, suppose that we aim to detect a 5% shift in the prevalence of malnutrition from baseline. Then, we can calculate $P(X_t \leq d|f_0(\cdot)+0.05)$ (see Section 3.8.4 for how to calculate this quantity). The advantage of using the Betabinomial model to select a survey design (as opposed to fixing p_l as a constant and using the binomial distribution), is that we accommodate uncertainty in the baseline rate of malnutrition. Without a substantial amount of contextual information, we anticipate that there will be uncertainty in the baseline rate of malnutrition across populations.

Scenario 3

Next, we discuss how to design a survey for Scenario 3, when we have limited historical information about the population of interest, along with data from a baseline survey. In

the baseline survey (at time $t-1$), we observed X_{t-1} malnourished children out of N_{t-1} children.

Before collecting data, we use historical information to estimate the baseline distribution of malnutrition in the population, e.g. $p \sim Beta(a_0, b_0)$. We specify the parameters a_0 and b_0 using prior knowledge of the population distribution of malnutrition. Choosing $a_0 = b_0 = 1$ assumes all prevalences are equally likely and may not be the optimal choice. If no historical information is available, we recommend selecting $a_0, b_0 < 1$. In our examples, we use $a_0 = b_0 = 0.1$. Poor specification of a_0 and b_0 can result in incorrect classifications.

To incorporate baseline data, the program manager should examine the data from the baseline survey and use contextual knowledge to determine if the survey results suggest malnutrition prevalence was high or low at baseline. If the data from the baseline survey is not consistent with the historical prior and prevalence appears high at baseline, then we should assume that prevalence is high at baseline and proceed with the monitoring and evaluation program accordingly. For instance, in the next survey following an intervention to attempt to lower malnutrition prevalence, we could aim to detect whether we have seen a drop in malnutrition prevalence from the baseline survey.

To design our survey, we assume malnutrition prevalence was relatively low and consistent with historical information at baseline. We aim to detect whether malnutrition prevalence increased from the baseline prevalence. Define $p_t - p_{t-1} = \Delta$. We choose upper and lower classification thresholds Δ_L, Δ_U based on the following criteria: if $\Delta > \Delta_U$, a notable rise in prevalence occurred from the previous time point; if $\Delta < \Delta_L$, no notable changes occurred. Typically, $\Delta_L = 0$ is a logical choice. Values between Δ_L, Δ_U are in the 'grey area,' and we do not restrict classification risks within the 'grey area' in the survey design.

Our objective is to find a minimum sample size at time t, N_t, and a decision rule d such that $Pr(X_t \leq d | \Delta = \Delta_U, X_{t-1}) \leq \alpha$ and $Pr(X_t > d | \Delta = \Delta_L, X_{t-1}) \leq \beta$. The classification risk α is the probability of classifying the change in prevalence as sufficiently low when

when $\Delta \geq \Delta_U$, and β is the probability of classifying the change in prevalence as high when $\Delta \leq \Delta_L$.

Assuming $X_{t-1} \sim Bin(N_{t-1}, p_{t-1})$, and given our prior knowledge a_0, b_0, we calculate $p_{t-1}|X_{t-1} \sim Beta(a, b)$, where $a = X_{t-1} + a_0$, $b = N_{t-1} - X_{t-1} + b_0$. We construct a distribution for $p_{t-1}|X_{t-1}$ and determine how likely it is that p_t was drawn from this distribution, shifted by Δ_L or by Δ_U.

Given N_t and d, we can calculate the OC probabilities $OC(\Delta) = P(X_t \leq d|\Delta, X_{t-1})$ at Δ_L and Δ_U (see Section 3.8.4), and obtain α and β for the design. We iterate through choices of N_t and d to find the optimal design that minimizes N_t and meets the specified classification risk thresholds.

Conceptually, this survey design is slightly different from those in Scenarios 1, 2, and 4. In this design, we compare the prevalence at time t to the prevalence at time $t-1$, accounting for uncertainty in the estimate of p_{t-1}. In the other designs, we compare the prevalence at time t to the baseline rate or baseline distribution of malnutrition.

Scenario 4 - comparing changes in prevalence over multiple time points

In Kenya and the Sudan, surveys are conducted every 6 months to monitor malnutrition. We aim to detect rises in the prevalence of malnutrition. In the previous section, we compared changes in prevalence from the previous time point (e.g. from baseline). Now, we consider how to combine information over multiple time points (e.g. a baseline, six-month, and one-year survey) to detect a rise in malnutrition.

As an example, consider comparing the one-year survey to the baseline and six-month survey. If the prevalence did not change between baseline and six-months, then we should pool the information across these two surveys and compare the pooled prevalence to the prevalence at the one-year survey. But, if the prevalence dropped between baseline and six-months, we aim to sustain this lower prevalence and thus compare the prevalence at one-year to the lower six-month prevalence. If the prevalence rose between

baseline and six months, then we determine if this rise is sustained at one-year. We denote the baseline, 6-month, and one-year prevalences as p_1, p_2, and p_3, respectively. To summarize, we aim to determine if p_3 is greater than the minimum of p_1 and p_2.

To detect rises in malnutrition prevalence, we can pool together information from the time points that are within ϵ of $min(p_1, ..., p_{t-1})$ to estimate $f_0(\cdot)$. For instance, we could pool together information from surveys that are within 2% of $min(p_1, ..., p_{t-1})$. In order to pool information between surveys, we directly employ the weighted method of moments estimator in Yousry et al. (1991), and empirically calculate $P(p_i < min(p_1, ..., p_{t-1}) + \epsilon)$ to obtain weights for the mean and variance estimators.

Choosing $\epsilon > 0$ results in pooling of more information, but is less 'conservative' than choosing $\epsilon = 0$. We could try to find an optimal ϵ for the survey design, but we suggest choosing either 0 or a value that is a fraction of Δ, e.g $1/5$ or $1/10$ of $\Delta_U - \Delta_L$. In our applications, we choose $\epsilon = (\Delta_U - \Delta_L)/5 = 2\%$, when $\Delta_L = 0$ and $\Delta_U = 10\%$.

Assuming that the baseline distribution $f_0(\cdot)$ follows a $Beta(a, b)$ distribution, we can then estimate a and b, using historical information and the past survey data. See Section 3.8.5 for more information on how to calculate these weights and estimate the baseline distribution $f_0(\cdot)$.

The advantage of using this surveillance tool is that we can pool together historical prior information and data from previous surveys. The approach is conservative, in that we compare the next time point to the "best case scenario," when prevalence was low. By using historical information, we avoid comparing the current time point to any extreme minima, by shifting the minimum toward the historical mean prevalence. If the prevalence has been chronically high throughout the surveillance program, then this design is not the best surveillance tool to use; malnutrition prevalence must be 'in control' during at least one surveillance time point to detect rises in prevalence. Further, given that the baseline distribution is estimated by collapsing information across surveys, our estimate of the baseline distribution may be inaccurate when we have data from very few surveys,

or there exists one survey that is an outlier.

3.4.2 Clustering in Temporal Surveys

In Section 3.4, we propose a surveillance tool for detecting rises in malnutrition prevalence over time, assuming the data was generated from a simple random sample. When data is collected using cluster sampling, we can use the design effect to adjust the sample size, as in Section 3.3.

Consider a survey comparing prevalence at time t to the prevalence at time $t-1$ (Scenario 3). First, we estimate ρ using the data from time $t-1$ and calculate the effective sample size for the survey at time $t-1$. Then, we update the distribution of $p_{t-1}|X_{t-1}$, $Beta(a,b)$, using the effective sample size rather than the original data, to incorporate the additional uncertainty in our sample due to clustering. To calculate the sample size for the survey at time t, we repeat the algorithm in Section 3.3, but perform the calculations in step 2 using the betabinomial distribution at time $t-1$, shifted by Δ_L and Δ_U, rather than the standard binomial model.

When adjusting for clustering with multiple surveys (Scenario 4), we assume ρ is constant over time and estimate across the surveys to obtain a stable estimate of ρ. Again, we would calculate the effective sample size for each of the previous surveys and construct $f_0(\cdot)$ by pooling information across the surveys using their effective sample sizes. Then, we again repeat the algorithm in Section 3.3, but perform the calculations in step 2 using the betabinomial distribution, $f_0(\cdot)$ shifted by Δ_L and Δ_U, rather than the standard binomial model.

3.5 Data application - survey designs in Kenya and South Sudan

3.5.1 Impact of clustering on the survey design

Using the 33×6 LQAS data from Kenya and South Sudan, we assess the impact of clustering on the survey design, using the methods presented in Section 3.3. We define malnutrition using the GAM indicator constructed using WHZ scores, and we assume that there are an infinite number of clusters in each survey site. This assumption is most likely violated, but we do not have information about how the clusters were enumerated.

For each survey, we estimate the intraclass correlation ρ using maximum likelihood estimation, assuming the data are generated from a betabinomial distribution. Ridout et al. (1999) summarizes many estimators of ρ for binary data. If the data do no follow a betabinomial distribution, our estimate of ρ may be poor. The average intracluster correlation over all of the surveys is $\hat{\rho} = 0.037$, and ranges from 0.00 to 0.13 across the surveys. Estimates of ρ using the betabinomial model were similar to the Ridout et al. (1999) ANOVA estimator, suggesting the betabinomial model is reasonable in this setting.

The estimate $\hat{\rho} = 0.037$ is a somewhat low, but non-negligible, intraclass correlation for binary data. For instance, in a 33×6 survey using the 15-20% couplet, if $\rho = 0$, the α and β risks are 0.14 and 0.22, respectively; if $\rho = 0.037$, these risks are now 0.16 and 0.24. These risks will be higher for the surveys with $\rho > 0.037$.

Advertising this survey design as having an α and β level of 0.1 and 0.2 is off the mark. Fixing $m = 6$, when $\rho = 0$, we would actually need to sample 44 (rather than 36) clusters to meet the desired risk levels. With $\rho = 0.037$, we would need to sample 53 (exact) or 57 (DEFF) clusters. If the number of clusters in the population was finite, say $K = 50$, then we then need to visit 46 clusters. For $\rho = .1$ with an infinite number of clusters, we would need to sample 66 (DEFF) or 67 (exact) clusters.

In Figure 3.5, we plot the relationship between m and k for the 15-20% couplet when

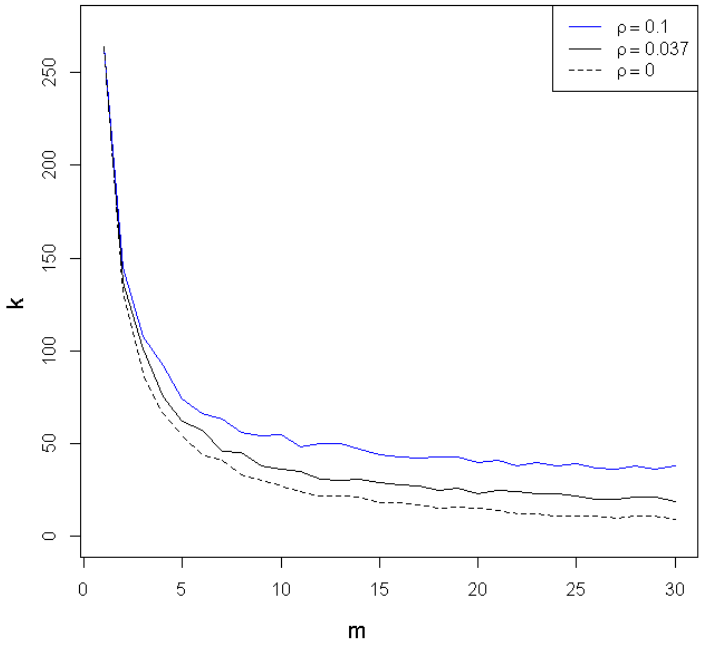


Figure 3.5: Relationship between k and m for $p_l = 0.15, p_u = 0.2, \alpha = 0.1, \beta = 0.2$.

$\rho = \{0, 0.037, 0.1\}$. As m gets large, k converges to a constant, illustrating the concept that the effective sample size plateaus as m increases. Further, for small m, k is similar between designs, because variance inflation due to clustering increases as m increases.

Sampling very few children per cluster may not be cost-effective, depending on the distance between clusters. Visiting 30-60 clusters/villages (depending on ρ) and only measuring 6 children per village could result in wasting both time and resources travelling between the villages. Rather than using cookbook designs, we can estimate the number of children that a field team could measure in one day, say $m = 20$. Then, we can use the LQAS sample size calculators, adjusting for intraclass correlation, to calculate the number of clusters we need to visit. When $\rho = 0.037$, using the design effect sample size formula, we need to travel to 23 clusters. So, now we would actually sample 430, rather than $53 * 6 = 318$ children if $m = 6$. However, the cost of the survey may decrease due to the reduction in travel expenses.

Table 3.3: Comparing time two to time one. Columns denote prevalence $\hat{p}_t$ and intraclass correlation coefficient $\hat{\rho}_t$ estimates at each time point; required sample sizes and decision rules for time two, X and d, and Δ?, denoting whether a rise in prevalence occurred at time two.

	Estimates				No Clustering				Clustering			
Loc.	$\hat{p}_1$	$\hat{p}_2$	$\hat{\rho}_1$	$\hat{\rho}_2$	N	d	X	Δ?	N	d	X	Δ?
GP	18.3	15.6	.05	.04	116	25	18.1	N	168	36	26.3	N
GR	12.5	20.7	006	.07	89	14	18.4	Y	126	20	26.1	Y
GU	17.0	11.5	.04	.05	131	27	15.1	N	156	31	18.0	N
MP	22.7	30.8	.01	.00	156	41	48.1	Y	168	44	51.8	Y
MR	36.6	38.3	.05	.00	248	100	-	?	432	174	-	?
MU	25.6	22.9	.00	.13	165	48	37.7	N	168	49	38.4	N
MS	20.7	4.3	.05	.03	136	33	5.8	N	192	46	8.2	N
SR	25.8	19.7	.03	.05	187	55	36.8	N	228	67	-	?
SU	20.0	22.4	.00	.04	136	32	30.4	N	138	32	30.9	N

3.5.2 Examining changes over time

We now use the survey designs presented in Section 3.4 to assess whether changes in GAM prevalence have occurred over time. First, we use the data at baseline, and compare the subsequent time point to determine if a spike in prevalence has occurred (Scenario 3 in Section 3.4). We plug-in the estimate of ρ from the previous survey to account for clustering in the surveys. We do not implement the designs to detect changes between multiple time points, due to the limited number of surveys (three per location) and insufficient historical information.

We choose $\Delta_L = 0, \Delta_U = 0.1$, $\alpha = 0.1$, and $\beta = 0.2$; that is, we aim to detect 10% changes in prevalence at least 90% of the time. We will accept detecting a change in prevalence when none has occurred at most 20% of the time. The survey designs and results are presented in Table 3.3. In our analysis, we have data from the 33×6 surveys, rather than from surveys with the recommended sample sizes. Therefore, given the observed data, we say that "a change has occurred over time" for a location if we would conclude that a change occurred the majority of the time if we randomly sampled from the collected survey data. For some of the surveys, we have not collected enough data

with the 33×6 design to reach a conclusion as to whether or not the prevalence changed over time.

In the Garissa Riverine and Mandera Pastoral populations, we detect a rise in prevalence over time from baseline. Using the highest 15-20% couplet, we would classify the Mandera Pastoral population as high prevalence at all three time points. Using the design to detect changes, we obtain more information - not only is the prevalence high in this region, it is on the rise.

3.6 Discussion

In this article, we discuss extensions to LQAS survey designs for monitoring malnutrition that improve the accuracy and flexibility of the existing designs. Using a simple design effect adjustment, LQAS surveys can be designed to preserve the prespecified classification risks when cluster sampling is used. Further, if the number of clusters in the population is finite, we adjust the sample size downward. Due to the potential impact of rounding errors, we caution against using the design effect adjustments in surveys with very small sample sizes (< 30) or with very rare events. In these scenarios, if the number of clusters is large, an exact method (such as Hedt-Gauthier et al. (2012)) is preferable.

Additionally, we present a cohesive surveillance tool for monitoring the prevalence of acute malnutrition over time. By combining historical information with data from previous surveys, we estimate the baseline distribution of "acceptable" malnutrition rates and detect spikes in malnutrition by comparing the collected data to this distribution. When the baseline distribution is iteratively updated, program managers must keep track of the following information: (1) number of malnourished children, (2) total sample size, (3) ρ, and (4) estimated design effect (if the sample size per cluster is constant over time and the number of clusters in the population is large, then tracking the design effect is not necessary). We anticipate that this longitudinal surveillance tool will be useful in any program aiming to detect deviations from a baseline rate, where the exact baseline rate is

unknown.

Using this survey design, we were able to detect rises in malnutrition prevalence in longitudinal programs in Kenya and the Sudan, where the baseline rates of malnutrition are known to vary.

3.7 Comparing Classical and Bayesian LQAS designs

LQAS survey designs for monitoring GAM have been criticized for the difficulty in interpreting the survey results and for producing too many false positives (Bilukha, 2008; Bilukha and Blanton, 2008). To address these criticisms, Olives and Pagano (2010) illustrate that using a Bayesian approach is necessary to control false positive and false negative rates. Myatt and Bennett (2008) provide another interesting public health application of a Bayesian classification procedures, using sequential survey designs for monitoring transmitted HIV drug resistance in developing countries.

In this section, we illustrate the properties of Bayesian and classical LQAS surveys. The classical LQAS classification procedure requires specification of the classification risks α and β. When $p \geq p_u$, α is the bound on the probability of classifying low; when $p \leq p_l$, β is the bound on the probability of classifying high. The conditional probabilities α and β condition on whether the true prevalence p in an area is $\geq p_u$ or $\leq p_l$, respectively.

The survey design procedure does not bound the classification risk between p_l and p_u (i.e. areas in the grey region). To calculate our sample size and decision rule, we only specify α, β, p_l, and p_u, and consequently do not specify whether it is an error to classify areas in the grey region as high or low ($p_l < p < p_u$). LQAS designs are constructed to ensure that areas with prevalences in the extremes (*i.e.* not in the grey region) are classified correctly.

To design a comparable Bayesian LQAS survey/classification procedure, we again specify upper and lower thresholds, p_l and p_u. We specify classification risks α_B and β_B,

with a different interpretation than α and β. In a Bayesian LQAS design, α_B is the probability that $p > p_u$, given that the prevalence in an area is classified as low; β_B is the probability that $p < p_l$ given that the prevalence in an area is classified as high. In the Bayesian design, the probabilities α_B and β_B are conditional on the classification decision (high or low.). In the design phase, we again make the implicit assumption that classification of areas in the grey region as either high or low is acceptable. In B-LQAS surveys, the length of the grey region can be set to 0.

We could design a Bayesian classification procedure based on different criteria than specifying α_B and β_B, *e.g.* using the figure of merit or specifying a loss function (Olives and Pagano, 2010). We discuss Bayesian survey designs based on the classification risks α_B and β_B to facilitate contrasting the Bayesian and classical survey designs. The discussion does not depend on which criteria are used to select a design.

The relationship between Bayesian and classical acceptance sampling designs is somewhat analogous to the relationship between sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). In the disease testing context, sensitivity is the probability that an individual tests positive given that he is disease positive; specificity is the probability that an individual tests negative given that he is disease negative. PPV is the probability that an individual who tests positive is disease positive; NPV is the probability that an individual who tests negative is disease negative. The difference between sensitivity, specificity, PPV, and NPV is the reversal of the conditioning event. Sensitivity and specificity condition on the disease status of an individual and are therefore properties of the test. PPV and NPV condition on the result of the test and consequently depend on the prevalence of the disease in the population.

Likewise, the fundamental difference between Bayesian and classical LQAS designs is the reversal of the conditioning event. In a classical LQAS design, α and β are calculated conditional on p and are therefore properties of the survey design. In a Bayesian design, α_B and β_B are calculated conditional on the classification decision (e.g. high or low) and depend on auxillary information, specifically a prior distribution of p.

Classical	Bayesian
$\alpha = P(\text{Classify low} \mid p \geq p_u)$	$\alpha_B = P(p \geq p_u \mid \text{Classify low})$
$\beta = P(\text{Classify high} \mid p \leq p_l)$	$\beta_B = P(p \leq p_l \mid \text{Classify high})$

Criticisms of classical LQAS designs often point out the high number of 'false positive' or 'false negative' classifications. However, we can calculate 'false positive' and 'false negative' rates only if we know the underlying population distribution for p, the prior distribution. Classical designs do not control for the number of false positive or false negative classifications; Bayesian designs do control the false positive and false negative rates, assuming the prior is known.

False positive and negative rates

To calculate the false positive rate, $P(p > p_u|\text{classify low})$, or false negative rate, $P(p < p_l|\text{classify high})$, from a survey, we need to specify the prior distribution for malnutrition prevalence. To perform these calculations, we could specify a non-informative prior for p, such as a $Beta(1, 1)$. Then, we could estimate the posterior distribution for p and the false positive and false negative rates, assuming p is equally likely to take on all values between 0 and 1. If we specify an informative prior distribution, we could estimate the posterior distribution for p more precisely. However, if we misspecify the informative prior, our estimate of p will be biased, along with the false positive and false negative rates. As the sample size increases, our data will dominate the prior information.

Similarly, we could use a non-informative prior to design a Bayesian acceptance sampling plan. For instance, if we specify a $Beta(1, 1)$ prior, the classification risks α_B and β_B are then false positive and false negative rates, *assuming p is equally likely to take on all values between 0 and 1*. When this assumption is not met, α_B and β_B are not true false positive and false negative rates for the population of interest. When a noninformative prior is used, decision rules and sample sizes will be similar to those from the classical design (Olives and Pagano, 2010).

If we incorporate more prior information, the classical and Bayesian sampling de-

signs will differ. If we misspecify the prior distribution, then the risks α_B and β_B will be biased estimates of the false positive and false negative rates. For instance, if our prior suggests malnutrition is high when prevalence is actually low, we will underestimate the false positive rate; if the prior suggests prevalence is low when it is truly high, we will overestimate the false negative rate, potentially missing malnutrition emergencies. Bayesian acceptance plans provide the minimum sample size required to satisfy the classification risks; consequently, the data may not dominate the prior distribution when an informative prior is selected.

Definition of the prior

To use a Bayesian acceptance sampling plan, the prior must be known. One can conceptualize the prior in the following manner: at a given location and time, the prevalence of malnutrition p is random variable, drawn from a known distribution - the prior. Bayesian designs have been used in manufacturing, because measuring the rate of random fluctuations in a machine over time allows construction of an accurate prior distribution; for a given machine, this prior is constant over time. Construction of this prior outside of the controlled manufacturing setting is a more difficult because this prior will change over time.

Consider the following hypothetical example. A population has a baseline rate of malnutrition between 3% and 7%; denote the baseline distribution of malnutrition prevalence as $f_0(\cdot)$. If a crisis occurs (e.g. a war or a drought), the prevalence of malnutrition spikes and is between 15% and 20%; denote the crisis distribution of malnutrition as $f_1(\cdot)$. Then, we can write the prior distribution as:

$$f(p) = w_0 f_0(p) + w_1 f_1(p)$$

for $p \in (0, 1)$, where w_1 is the probability that a crisis has occurred and malnutrition prevalence has spiked, and $w_0 = 1 - w_1$. To specify a prior for a Bayesian design, we would need to know 3 quantities: $f_0(\cdot), f_1(\cdot)$, and w_0. The distributions of $f_0(\cdot)$ and $f_1(\cdot)$ can be estimated from historical data or knowledge of the program managers, when available.

To specify w_0, we need to know the likelihood that we are in a crisis setting, *i.e.* the probability that the malnutrition prevalence is high. However, this is presumably the goal of the survey - to detect whether a rise in malnutrition has occurred.

Classical LQAS survey designs always maintain the specified classification risks α and β, regardless of the underlying population distribution. The cost of using the classical design is that we are often more interested in the Bayesian interpretation, *i.e.* false positive and false negative rates (Bilukha and Blanton, 2008; Olives and Pagano, 2010). However, for a Bayesian survey design, whether or not the classification risks α_B and β_B are correct depends completely on the correct specification of the prior distribution. The design parameters α_B and β_B control the false positive and false negative rate, but are only interpretable with respect to the prior that was selected. If the prior is misspecified, α_B and β_B are no longer interpretable.

3.8 Statistical derivations of the survey design attributes

We have proposed numerous adjustments to the LQAS survey designs for monitoring malnutrition. In this section, we derive statistical properties of these adjustments and describe how to perform the subsequent calculations.

3.8.1 Derivation of design effect formula

First, we define the necessary notation. The population contains K clusters (PSUs), and k are sampled; within each clusters, there are M individuals (SSUs), and m are sampled. The survey weights are equal for all individuals in the population, because **we are assuming that clusters are approximately the same size**. The intraclass correlation ρ is defined as:

$$\rho = 1 - \frac{M}{M-1}\frac{\sigma_W^2}{\sigma_W^2 + \sigma_B^2}$$

For large M, $\rho = 1 - \frac{\sigma_W^2}{\sigma_W^2+\sigma_B^2}$. Then, as $K \to \infty$,

$$\begin{aligned} DEFF &= \frac{Var_{clus}(\hat{p})}{Var_{SRS}(\hat{p})} \\ &= \frac{(1-\frac{k}{K})\frac{\sigma_B^2}{k} + (1-\frac{m}{M})\frac{\sigma_W^2}{km}}{(1-\frac{km}{KM})\frac{\left(\sigma_B^2+\sigma_W^2\right)}{km}} \\ &= \frac{m(1-\frac{k}{K})\rho}{(1-\frac{km}{KM})} + \frac{(1-\frac{m}{M})(1-\rho)}{(1-\frac{km}{KM})} \\ &\approx m(1-\tfrac{k}{K})\rho + (1-\rho) \\ &\approx 1 + (mf-1)\rho \end{aligned}$$

where $f = (1-\frac{k}{K})$. Note that when the number of clusters K is large, $DEFF \approx 1 + (m-1)\rho$.

3.8.2 Derivation of effective sample size asymptote

Denote the effective sample size as $k \to \infty$ as n^*.

$$\begin{aligned} n^* &= \frac{n}{DEFF} \\ &= \frac{n}{1+(mf-1)\rho} \\ &= \frac{km}{1+(m\{1-\frac{k}{K}\}-1)\rho} \\ &= \frac{k}{\frac{1}{m}+(1-\frac{k}{K}-\frac{1}{m})\rho} \\ &= \frac{k}{(1-\frac{k}{K})\rho} \end{aligned}$$

3.8.3 Moment estimators for the Beta distribution

The mean and variance of a Beta(a,b) distribution are $\mu = a/(a+b)$, $V = (ab)/(\{a+b\}^2\{a+b+1\})$. Let $\theta = (1-\mu)/\mu$. Then, $a = \theta/(V\{1+\theta\}^3) - 1/(1+\theta)$. $b = a\theta$.

Given data from T distinct time points, Yousry et al. (1991) suggests estimating

$E(X_t/n_t)$ and $Var(X_t/n_t)$ as

$$\hat{\mu} = \frac{\sum_{t=1}^{T} \lambda_t X_t}{\sum_{t=1}^{T} \lambda_t n_t} \text{ and } \hat{V} = \frac{\sum_{t=1}^{T} (\lambda_t n_t \{X_t/n_t - \hat{\mu}\}^2)}{\sum_{t=1}^{T} \lambda_t n_t}.$$

where $0 < \lambda_t < 1$ are weights. Recursive forms of these equations are provided in Yousry et al. (1991) for fast updating.

To incorporate historical information, we estimate $E(X_t/n_t)$ and $Var(X_t/n_t)$ as:

$$\hat{\mu} = \frac{\sum_{t=1}^{T} \lambda_t (X_t + a_0)}{\sum_{t=1}^{T} \lambda_t (n_t + a_0 + b_0)} \text{ and } \hat{V} = \frac{\sum_{t=1}^{T} (\lambda_t (n_t + a_0 + b_0) \{(X_t + a_0)/n_t - \hat{\mu}\}^2)}{\sum_{t=1}^{T} \lambda_t (n_t + a_0 + b_0)}.$$

Consequently, we smooth the individual estimates at each time point toward the historical data, diminishing the impact of outliers and reducing the variance of the baseline distribution as we include more historical information.

3.8.4 Evaluating $Pr(X_t \leq d|\Delta, X_{t-1})$

Denoting the density for p_t by $f_0(\cdot)$, we calculate

$$Pr(X_t \leq d|\Delta, X_{t-1}) \quad = \quad \int P(X_t \leq d|p, \Delta, X_{t-1}) f_0(p) dp$$

noting that $p_t \sim Beta(a, b)$, and $X_t|p_t, \Delta \sim Bin(N_t, p_t + \Delta)$. This integral can be evaluated using MCMC integration.

Alternatively, to obtain a closed form estimate of this integral, we can assume that $p_t \sim Beta(a, b)$, where a and b are estimated using the method of moments as follows:

1. Calculate $\hat{\mu}_{t-1}$ and $\hat{V}_{t-1}$, the estimated mean and variance of the $Beta(a, b)$ distribution.

2. Assume $\hat{\mu}_t = \hat{\mu}_{t-1} + \Delta$, $\hat{V}_t = \hat{V}_{t-1}$, where $\hat{\mu}_t$ and $\hat{V}_t$ are the mean and variance of the $Beta(a, b)$ distribution, respectively.

3. Calculate a and b using the moments $\hat{\mu}_t$ and $\hat{V}_t$.

Now, we can obtain a closed form expression for the OC probabilities by using the betabinomial model for X_t, specifically $X_t \sim Betabinomial(a, b)$.

$$Pr(X_t \leq d|\Delta, X_{t-1}) = \sum_{i=0}^{d} \left(\begin{array}{c} N_t \\ i \end{array} \right) \frac{B(i+a, N_t - i + b)}{B(a, b)}$$

where $B()$ is the beta function.

The MCMC integration and MOM approaches will provide similar results for sufficiently large a, b, due to the approximate asymptotic normality of the $Beta$ distribution. Conceptually, the MCMC distribution compares X_t/N_t to a shifted $Beta$ distribution, whereas the method of moments defines a new $Beta$ distribution by fixing the variance at time $t-1$ and shifting the mean by Δ.

3.8.5 Evaluating $P(X_t \leq d|\Delta, X_1, ..., X_{t-1})$

Denote X_i, N_i as the number of malnourished children and total sample size at time $i = \{1, ..., t\}$. We compare the prevalence of malnutrition at time t to the distribution of malnutrition at the previous time points, *when the prevalence was low*. To achieve this goal, we need to first construct a distribution that reflects the prevalence of malnutrition *when low*, over the previous $t-1$ time points, which we denote $f_0(\cdot)$. We assume that $f_0(\cdot)$ is a $Beta(a, b)$ distribution, where a, b are calculated using the weighted method of moments estimator. For $i = \{1, ..., t-1\}$, the weights are defined as $\lambda_i = P(p_i < min(p_1, ..., p_{t-1}) + \epsilon)$, where ϵ is a user-defined parameter. There is not a closed form for calculating the weights, and therefore this calculation is performed empirically, by sampling from the posterior distributions of $\{p_t\}$.

$$\begin{aligned} Pr(X_t \leq d|\Delta, X_1, ..., X_t) &= \int P(X_t \leq d|p, \Delta, X_{t-1}) f_0(p) dp \\ &= \int \sum_{i=0}^{d} P(X_t = i|p, \Delta, X_{t-1}) f_0(p) dp \end{aligned}$$

We evaluate this integral in the exact same way as in the above section, using either MCMC integration or method of moments.

A geostatistical approach to large-scale disease mapping with temporal misalignment

Lauren Hund[1], Jarvis Chen[2], Nancy Krieger[2], and Brent Coull[1]

[1]Department of Biostatistics
Harvard School of Public Health

[2]Department of Society, Human Development, and Health
Harvard School of Public Health

4.1 Introduction

Area-level aggregated count data arise frequently in the disease mapping setting (Best et al., 2005; Wakefield, 2007); for instance, in this paper, we assess the impact of socioeconomic disparities on breast cancer incidence by linking census data from multiple time points to cancer registry data. Our dataset is large ($\sim$ 2000 areas at each time point) and contains temporally misaligned boundaries, because census tract boundaries change over time. These types of data are becoming increasingly common in practice, due to our ability to merge census data and data from other large databases, such as disease registries.

Area-level data is most frequently modeled using hierarchical Bayesian models, with spatial correlation between areas incorporated through area-specific random effects having conditional Markov random field (MRF) priors (Besag et al., 1991). In the spatio-temporal setting where boundaries change over time, the use of area-specific random effects is not applicable because the areas are not well-defined over the course of the study (Chen et al., 2008).

To our knowledge, few approaches to spatial regression exist that allow for temporal boundary misalignment. Mugglin et al. (2000) and Zhu et al. (2000) address the boundary misalignment issue using hierarchical Bayesian models, with conditional Markov random field (MRF) priors on the area-specific random effects. These existing methods are computationally intensive, and typically bog down for large datasets. Zhu et al. (2000) suggest including time- and area-specific random effects in the linear predictor of the model, implicitly assuming that the area-level random effects are independent across time points. Because this independence assumption is typically violated in standard longitudinal settings, the resulting inferences on changes over time can be inefficient.

In this paper, we propose a geostatistical disease mapping model that allows for spatially misaligned boundaries over time. We model the underlying spatial continuous risk surface as a Gaussian random field (Kelsall and Wakefield, 2002; Best et al., 2000; Muller et al., 1997). We reduce the computational burden of spatial smoothing by modeling spa-

tial correlation using bivariate low-rank, penalized-splines (Kammann and Wand, 2003; Ruppert et al., 2003). Area-level data is sometimes treated as point-referenced based on the centroid of an area, and the penalized-spline/mixed model approach is then used to model spatial variability, *e.g.* Lee and Durban (2009). However, these models also do not directly incorporate information about the size and shape of each area and can perform poorly (Best et al., 2005). By modeling the underlying spatial risk surface and aggregating to the area-level, we overcome these limitations.

We implement the model within the generalized linear mixed model (GLMM) framework, modeling the underlying spatial surface using radial basis splines (Kammann and Wand, 2003; Ruppert et al., 2003), facilitating fitting a reduced-rank, computationally fast version of the model. We estimate model parameters using a penalized quasi-likelihood approximation to maximum likelihood estimation (Breslow and Clayton, 1993). Similar to the Kelsall and Wakefield model, our approach has the desirable property that smaller areas have larger prior variances. Additionally, the actual shape of each area, as opposed to only the neighborhood structure, is incorporated into the covariance between areas, avoiding any problems that could arise from oddly-shaped areas. Our method is easy to program in standard statistical software packages and is not computationally intensive relative to MRF formulations.

Section 4.2 introduces the motivating study for our methodology. Section 4.3 describes the formulation of the geostatistical disease mapping model, and Section 4.4 presents spatio-temporal extensions of the model. Sections 4.5 and 4.6 give the results of a simulation study and data analysis, respectively.

4.2 Motivating study: Breast cancer incidence in Los Angeles

Breast cancer is presently the leading cause of cancer among U.S. women (excluding non-melanoma skin cancers), accounting for 28% of the diagnosed cases (American Cancer

Society, 2010). Breast cancer typically has been portrayed as a "disease of affluence." As secular changes in the socioeconomic distribution of breast cancer risk factors occur, incidence rates in poorer countries and among poorer women in more affluent countries may be "catching up" over the long term (Krieger et al., 2006). Examining the changes in the socioeconomic distribution of breast cancer incidence is important for public perception and policies regarding the disease, as well as to gauge the population mortality burden of breast cancer (Krieger et al., 2006). We investigate the hypothesis that the socioeconomic gradient in breast cancer incidence is decreasing over time by examining data associations between socioeconomic measures and breast cancer incidence rates across two decades.

We apply our method to assess changes in the socioeconomic gradient of breast cancer in women over time in Los Angeles County, CA, focusing on the time periods 1988-1992 and 1998-2002, which precedes the change in breast cancer incidence rate attributed to declining use of hormone therapy. Krieger et al. (2006) originally analyzed these data by calculating age-standardized breast cancer incidence rates, stratified by decade, race/ethnicity, and socio-economic status, and ignoring spatio-temporal correlation between areas. Our analysis parallels this original report, but incorporates spatio-temporal information into a regression model, yielding a more efficient analysis.

We quantify the socioeconomic gradient by calculating the difference in the breast cancer log-incidence rate ratios corresponding to an area-based socioeconomic measure (ABSM) for the time periods 1988-1992 and 1998-2002. We obtain total population counts of women by age and race/ethnicity and poverty indicators (ABSMs) from U.S. census data at the census tract (CT) level in L.A. county for 1990 and 2000. There are a total of 1,642 census tracts in 1990 and 2,056 census tracts in 2000, reflecting a large number of census tract boundary changes between the two time periods. We obtained the breast cancer case data from the Los Angeles Cancer Surveillance Program cancer registry. We appended the census tract geocode to each cancer registry record, based on the location and date of residence at diagnosis. We link incident cases between 1988-1992 to the 1990 census population data and cases between 1998-2002 to the 2000 census data.

Figure 4.1: Examples of changes to census tract boundaries from one census to the next.

U.S. census tract boundaries are redefined over time as necessary to maintain an average population between 3,000 and 4,000 in each census tract, with each tract relatively socioeconomically homogeneous (US Census Bureau, 1994). Figure 4.1 illustrates different types of changes in census tract boundaries. Because changes do not always take the form of simple splitting or merging, there is not a one-to-one correspondence between census tracts over time.

4.3 Statistical framework for the spatial model

Best et al. (2005) and Wakefield (2007) review standard spatial disease mapping models. We use the following notation for our disease mapping model. Observed cases of a disease Y_i within an area A_i are modeled using a Poisson likelihood, $Y_i \sim \text{Poisson}(e^{S_i} E_i)$, for $i = 1, ..., M$. We model S_i, the log-relative risk of disease, as a function of covariates and spatial random effects. Assuming disease prevalence varies within certain strata j (such as age groups), we calculate the expected number of cases in a region E_i using the prevalence of disease and the population count in each strata (i.e. using internal or external standardization). Our model assumes that the disease is rare and that the risk associated

with living in area i acts proportionally on the baseline risks for each stratum.

We now develop a geostatistical model for spatial correlation that is similar in nature to Kelsall and Wakefield (2002), which we refer to as KW throughout this paper. Consider an area A which is partitioned into regions $\{A_i\}$. Specifically, let *i* index census tracts (CTs) in region A, $i = 1, ..., M$, and s_{ij} be a point location within A_i, $s_{ij} \in A_i$. Define $|A_i|$ as the area of A_i; Y_i as the number of events in A_i; $\lambda(s)$ as the intensity of the Poisson process at point s; and $f_i(s)$ as the population density in A_i at point s. If we assume the population density is uniform over A_i, then $f_i(s) = 1/|A_i|$. If more information is available about the population density within an area A_i, we can use a piecewise uniform surface to estimate $f_i(s)$.

The diseased cases $Y(s)$ follow a Poisson process with intensity $E_i f_i(s) R(s)$, where $R(s)$ is the relative risk of disease at location s. Aggregating to the area-level, $Y_i \sim \text{Poisson}\{E_i \int_{A_i} f_i(s) R(s) ds\}$, and the average relative risk in area A_i is $R_i = \int_{A_i} f_i(s) R(s) ds$. Disregarding spatial and covariate effects, $Y_i \sim \text{Poisson}(E_i R_i)$.

We incorporate covariates and spatial random effects through modeling the log-relative risk as $S(s) = \log R(s) = S'(s) + \beta Z(s)$, where $Z(s)$ is the covariate surface and $S'(s)$ is a continuous surface inducing spatial correlation between areas. KW propose a multivariate normal model for the area-level log-relative risk, with the covariance between the two areas interpreted as the average covariance between two points chosen randomly from the two areas. Diverging from KW, we model $S'(s)$ using a penalized spline term.

4.3.1 Approximating the log-relative risk

We construct our model as a generalized linear mixed model (GLMM) using radial splines to model spatial correlation (Ruppert et al., 2003). The underlying model for the log-relative risk is $S(s) = \beta X(s) + S'(s) = \beta X(s) + \sum_l Z_l(s) u_l$, where we write the spatial terms of the model $S'(s)$ as a penalized spline term. The basis functions $\{Z_l(s)\}$ are known, de-

rived from a set of knots on the area A and a standard spatial covariance function (which we discuss in Section 4.3.2), and the $\{u_l\}$ terms are basis coefficients assumed to be independent normal random effects estimated via model fitting. By using this penalized spline representation, we express the model for the underlying relative risk as a generalized linear mixed model (GLMM). We use a quadrature approximation to estimate the spatial random effects for each area:

$$S_i = \int_{A_i} f_i(s) \left\{ X(s)\beta + \sum_l Z_l(s)u_l \right\} ds \approx X_i\beta + \sum_l \sum_j w_{ij} Z_l(s_{ij})u_l.$$

where $\{s_{ij}\}_{j=1,...,d_i}$ are the d_i design points selected area A_i, and $\{w_{ij}\}$ are the corresponding quadrature weights for each area ($\sum_j w_{ij} = 1$ where $j = 1, ..., d_i$). If $\{X_i\}$ is an aggregate-level covariate, then $X_i = \int_{A_i} X(s) f_i(s) ds$ whereas if $\{X_i\}$ is inheritable, $X_i = X(s)\ \forall s \in A_i$. Conclusions drawn based on inheritable covariates are subject to ecological bias if we extrapolate area-level results to individuals (Wakefield, 2007).

In order to fit our model, appropriate design points $\{s_{ij}\}$ and corresponding quadrature weights $\{w_{ij}\}$ must be selected. If the design points correspond to sub-areas with known population counts (such as the centroids of block groups within census tracts), quadrature weights could be chosen to reflect the underlying population density, $w_{ij} = N_{ij}/N_i$, where N_{ij} is the total population size in sub-area A_{ij} and $N_i = \sum_j N_{ij}$ is the total population size in area A_i. Alternatively, assuming the population density is constant within an area, the best choice of design points corresponds to a grid of equally spaced points within each area; the resulting quadrature weights are $w_{ij} = 1/d_i$. With only one design point (the centroid) per area, our approach reduces to the model in which areas are treated as point-referenced data based on the centroid of the area, and a standard covariance function is specified to model spatial correlation between centroids.

4.3.2 Defining the spatial correlation structure

We write the *underlying* model for the log-relative risk in mixed model form as $\boldsymbol{S}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}^*\boldsymbol{u}$, where $\boldsymbol{S}^* = \{S_{ij}\}_{j=1,...,d_i; i=1,...,M}$; $\tilde{\boldsymbol{Z}}^*$ is a contrast matrix described below;

and $\boldsymbol{u} \sim \text{MVN}(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I})$. S_{ij} is the log-relative risk at design point j in area i. We construct $\tilde{\boldsymbol{Z}}^*$ such that $\text{Cov}(S_{ij}, S_{i'j'}) = C(|s_{ij} - s_{i'j'}|; \boldsymbol{\rho})$, where C is a standard spatial covariance function that depends on the distance between the design points and parameters $\boldsymbol{\rho}$. Due to the aggregate nature of the data, choice of a spatial covariance function is less important in this setting, as the model is not as sensitive to misspecification of the correlation function (see Section 4.5.2). We recommend using the exponential covariance function, $\text{Cov}(S_{ij}, S_{i'j'}) = \sigma^2 \text{exp}(-|s_{ij} - s_{i'j'}|/\rho)$ for its simplicity. We choose a value for the range parameter ρ by selecting a plausible value based on the fact that $3/\rho$ is the approximate distance at which the correlation between S_{ij} and $S_{i'j'}$ is less than 0.05 (Banerjee et al., 2003). Alternatively, we could select ρ by choosing a value that minimizes the model deviance.

We fit a reduced rank approximation of the model by choosing a set of knots $\{\kappa_g\}_{g=1,...,G}$ and basing our spatial correlation structure on the distances between the design points s_{ij} and the G knots (Kammann and Wand, 2003). When computationally feasible, we define the knots as the centroids of the areas in the study (G equals the number of areas in the study). A more practical approach is to use a knot selection algorithm to choose G knots in the study region, *e.g.* Johnson et al. (1990), which performs well in practice (Wand, 2003).

Define the $d_i \times G$ matrix $\boldsymbol{Z_i} = \{C(|s_{ij} - \kappa_1|), ..., C(|s_{ij} - \kappa_G|)\}_{j=1,...,d_i}$, which corresponds to the covariance between the design points in area i and the G knots. We stack the area-specific $\boldsymbol{Z_i}$ matrices to construct $\boldsymbol{Z} = (\boldsymbol{Z_i})_{i=1,...,M}$. Define the $G \times G$ matrix representing the covariance between the knots as $\boldsymbol{\Omega} = \{C(|\kappa_{g_1} - \kappa_{g_2}|)\}_{g_1,g_2=1,...,G}$. Then, $\tilde{\boldsymbol{Z}}^* = \boldsymbol{Z\Omega^{-1/2}}$. From the definition of $\tilde{\boldsymbol{Z}}^*$, it follows that $\text{Var}(\boldsymbol{S^*}) \approx \sigma_u^2 \tilde{\boldsymbol{Z}}^* \tilde{\boldsymbol{Z}}^{*\boldsymbol{T}} = \sigma_u^2 \boldsymbol{Z\Omega^{-1}Z^T}$.

Now, let S_i be the area level log-relative risk for area A_i. For $\boldsymbol{S} = (S_1, S_2, ..., S_M)$, $\boldsymbol{S} = \boldsymbol{X\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{u}$, where $\tilde{\boldsymbol{Z}} = \boldsymbol{W}\tilde{\boldsymbol{Z}}^*$, $\boldsymbol{X} = \boldsymbol{WX^*}$, and $\boldsymbol{W}$ is a $M \times \sum_{i=1}^{M} d_i$ block-diagonal matrix of the quadrature weights. Specifically, row i of $\boldsymbol{W}$ contains the d_i quadrature weights w_{ij} in the columns corresponding to area S_i and 0s everywhere else. Then, $\text{Var}(\boldsymbol{S}) = \sigma_u^2 \tilde{\boldsymbol{Z}} \tilde{\boldsymbol{Z}}^{\boldsymbol{T}} = \sigma_u^2 \boldsymbol{WZ\Omega^{-1}Z^TW^T}$. Since $\boldsymbol{S^*} \sim \text{MVN}\{\boldsymbol{X^*\beta}, \text{Var}(\boldsymbol{S^*})\}$ and $\boldsymbol{S}$ is a linear transformation of

$\boldsymbol{S}^*$, it follows that $\boldsymbol{S} \sim \text{MVN}\{\boldsymbol{X\beta}, \text{Var}(\boldsymbol{S})\}$. Examining the covariance between individual areas clarifies that the covariance matrix in our model has the same interpretation as that in the KW model. The covariance between the log-relative risk for areas A_i and A_j is $\text{Cov}(S_i, S_j) = \sum_{k=1}^{d_i} \sum_{l=1}^{d_j} w_{ik} w_{jl} \sigma_{kl} = \sum_{k=1}^{d_i} w_{ik} \sum_{l=1}^{d_j} w_{jl} \sigma_{kl}$, where σ_{kl} is the covariance between the log-relative risk at points s_{ik} and s_{jl}. That is, the covariance between two areas is a weighted average of the covariance between the design points in the area, and the variance of an area is the average covariance between the design points within an area.

4.3.3 Generalized linear mixed model construction

Using the above formulation of the log-relative risk within an area, we write the model as $Y_i \sim \text{Poisson}(e^{S_i} E_i)$, where $\boldsymbol{S} = \boldsymbol{X\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{u}$. In this model, $\boldsymbol{\beta}$ are fixed effect parameters, $\boldsymbol{S} = (S_1, ..., S_M)^T$, $\boldsymbol{X} = (x_1, ..., x_M)^T$, $\mathbf{u} = (u_1, ..., u_G)^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ are independent random effects, and $\tilde{\boldsymbol{Z}}$ is the $M \times G$ matrix defined in Section 4.3.2.

We introduce an overdispersion parameter ϕ into the model to account for additional non-spatial variability in the data greater than that predicted by the Poisson distribution. We consider other methods for incorporating residual overdispersion below and compare the performance of these various approaches 4.5.2. The likelihood for the model is:

$$L(\beta, \phi, \sigma^2; y_i) \propto (\sigma\sqrt{2\pi})^{-K} \int_{R^G} \exp\left\{ \sum_{i=1}^{M} \tfrac{1}{\phi}(-e^{\eta_i} + y_i \eta_i) + \sum_{l=1}^{G} -u_l^2/2\sigma^2 \right\} d\mathbf{u}.$$

The likelihood involves a G-dimensional integral, which is computationally expensive to evaluate. We approximate this integral using penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993). We have constructed an R package for fitting this model, available for download at http://www.hsph.harvard.edu/statinformatics/soft/areaglmm.html, and SAS code is available from the authors upon request.

In Section 4.5, we evaluate the performance of the PQL approximation for our model. Fitting our model using a Bayesian framework for the estimation of parameters is rel-

atively straightforward (Crainiceanu et al., 2005), though much more computationally intensive.

Modeling Overdispersion

We incorporate spatial random effects into our model that allow for spatially structured extra-Poisson variability. If residual non-spatially structured variability arises, we can incorporate this overdispersion in the regression model in several different ways. In Section 3.3, we suggest estimating an overdispersion parameter ϕ to account for nonspatial variability. This method uses quasi-likelihood estimation, specifying the mean and variance of Y_i ($E(Y_i|X_i) = \mu_i, \mathrm{Var}(Y_i|X_i) = \phi\mu_i$), but not the full distribution of Y_i. We call this the "quasi-Poisson" model. Alternatively, we could model Y_i using negative binomial regression to account for residual overdispersion by assuming that $Y_i|X_i, \psi_i \sim \mathrm{Poisson}(\psi_i\mu_i)$, and ψ_i follows a Gamma distribution, and then marginalizing over ψ_i. Lastly, we could add a normally-distributed random-intercept into the linear predictor of our model.

4.3.4 Mapping the relative risk surface

Constructing the smoothed predicted continuous relative risk surface or the smoothed predicted area-level relative risk surface is relatively straightforward. The pointwise relative risk estimates are $R(s) = \exp\left\{X(s)\hat{\beta} + \sum_l Z_l(s)\hat{u}_l\right\}$, and the area-level relative risk estimates are $R_i \approx \exp\left[\sum_j w_{ij}\left\{X(s_{ij})\hat{\beta} + \sum_l Z_l(s_{ij})\hat{u}_l\right\}\right]$, where the s_{ij} are design points in A_i with corresponding quadrature weights w_{ij}.

In order to obtain confidence bounds for the area-specific relative risk estimates, we estimate the standard errors for the fixed and random effects based on the PQL procedure (Ruppert et al., 2003). Specifically, $\mathrm{Cov}\left\{\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{u}}\end{pmatrix}|\boldsymbol{u}\right\} \simeq (\boldsymbol{C}^T\boldsymbol{V}\boldsymbol{C} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{C}^T\boldsymbol{V}\boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{V}\boldsymbol{C} + \sigma^2\boldsymbol{I})^{-1}$, where $\boldsymbol{C} = (\boldsymbol{X} \quad \boldsymbol{Z})$ and $\boldsymbol{V} = \mathrm{Var}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{u}) = \boldsymbol{e}^{\boldsymbol{S}}\boldsymbol{E}$ and $\boldsymbol{E}$ is the expected count in each area. The standard error of the linear predictor $\boldsymbol{S}$ is $\sqrt{\boldsymbol{C}\mathrm{Cov}\left\{\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{u}}\end{pmatrix}|\boldsymbol{u}\right\}\boldsymbol{C}^T}$, which we can use to obtain pointwise or area-specific confidence intervals.

4.4 Spatio-temporal extensions

Extending the model from the spatial to the spatio-temporal setting is straightforward using the spatial mapping methods of Wager et al. (2004). Assume we observe counts in M_t areas at time points $t = 1, ..., T$, where $Y_{it} \sim \text{Poisson}(E_{it}R_{it})$ and $i = 1, ..., M_t$. For notational simplicity, we assume knot locations are the same across time points, though this assumption is not necessary.

We propose three different spatio-temporal models for the underlying log-relative risk. First, if the underlying risk surface is the same shape at each time point and shifts by a constant over time, then we model the log-relative risk as (**Model 1**):

$$S_{ijt} = X_{ijt}\beta + \sum_l Z_l(s_{ijt})u_l + \delta_t,$$

where $u_l \sim N(0, \sigma^2)$ and δ_t is an intercept for time t. We note that $\delta_t = \delta t$, or some variation of this, might be more appropriate in some applications. Model 1 assumes perfect correlation between spatial random effects across time.

Another option for modeling spatial structure is to fit a model analogous to Zhu et al. (2000), where we do not allow for a common spatial surface across time points (**Model 2**):

$$S_{ijt} = X_{ijt}\beta + \delta_t + \sum_l Z_{lt}(s_{ijt})u_{lt},$$

where $u_{lt} \sim N(0, \sigma_t^2)$. In Model 2, spatial random effects are independent across time points, ignoring temporal correlation in the spatial random effects. $Z_{lt}(\cdot)$ are time-specific spline terms defined in the same way as in Section 4.3.2, except that the range parameter (or correlation structure) does not necessarily have to be equal across time points.

By adding an additional time-specific spatial random effect, we can fit a more flexible spatio-temporal model (**Model 3**):

$$S_{ijt} = X_{ijt}\beta + \delta_t + \sum_l \{Z_l(s_{ijt})u_l + Z_{lt}(s_{ijt})u_{lt}\},$$

where $u_l \sim N(0, \sigma^2)$ and $u_{lt} \sim N(0, \sigma_t^2)$ are independent random effects. In Model 3, the spatial relative risk surface differs at each time point due to the inclusion of the $\{u_{lt}\}$ ran-

dom effect terms. The shared spatial surface represented by the $\{u_l\}$s, which are constant across time, induce temporal correlation in the random effects. Unless data are extremely sparse, we recommend using Model 3 in practice.

The area-level model for the log-relative risk at time t is $\boldsymbol{S_t} = \boldsymbol{W_t S_t^*}$, where $\boldsymbol{S_t^*} = \{S_{ijt}\}_{i=1,...,d_i;j=1,...,M_t}$. $\boldsymbol{W_t}$ is a *time-specific* quadrature weight matrix, identical to $\boldsymbol{W}$, but specific to time t. For instance, the area-level log-relative risk for Model 3 is $\boldsymbol{S_t} = \boldsymbol{X_t \beta_t} + \boldsymbol{W_t \tilde{Z}^*(u + u_t)}$, where $\boldsymbol{u_t} = (u_{1t}, ..., u_{Gt})^T$ and $\boldsymbol{u_t} \sim \text{MVN}(\boldsymbol{0}, \sigma_{ut}^2 \boldsymbol{I})$. Then, $\text{Var}(\boldsymbol{S_t}) = \boldsymbol{W_t \tilde{Z}^* (\Omega_0^{-1} + \Omega_t^{-1}) \tilde{Z}^{*T} W_t^T}$, similar to the variance of the log-relative risk at a single time point.

Within this geostatistical framework, boundary misalignment between areas over time no longer requires complicated model fitting schemes. The quadrature weight matrix $\boldsymbol{W_t}$ is different between time points when boundaries are misaligned, because design points may lie in different areas across time as boundaries change. To understand how our method accounts for boundary misalignment, it is useful to think of the locations of the design points and knots that induce the underlying risk surface as being fixed across time (though this is not necessary in model fitting). Because we model the underlying risk surface through these reference design points and knots, changing boundaries are no longer problematic.

4.5 Simulation Study

We conduct a simulation study to assess performance of the model. Goals of the simulation study include: (1) quantifying gains in power for detecting changes in a covariate effect over time when we account for temporal correlation in spatial random effects; (2) confirming that bias is negligible in the fixed effects and variance components when we use the PQL approximation with sparse data, and (3) examining sensitivity of the model to choice of the range parameter ρ and to the number of knots used.

4.5.1 Design of simulation study

We briefly describe the design of our simulation study, but relegate the specific details to Section 4.8.

To construct our datasets, we start with a $(0,1) \times (0,1)$ regular grid of equal size areas, but relax this assumption shortly. We divide the grid into 64, 256, or 1024 square blocks (areas). We fix the disease incidence p at 0.11 cases per 100 person-years and the total population in the area at 9.5 million, similar to our data application. We define the expected number of cases in an area as the product of the disease incidence and the total population in the area.

We simulate a Poisson process with intensity λ_{ijt} at location j in area i at time t, where the log-intensity is (analogous to Model 3 in Section 4.4) $\log(\lambda_{ijt}) = \log(E_{ijt}) + \beta_x x_{it} + \beta_t t + \beta_{xt} x_{it} t + \xi(s_{ijt}) + \xi_t(s_{ijt})$. E_{ijt} is the expected number of cases at time t in area i at point j. $\xi(s_{ijt})$ and $\xi_t(s_{ijt})$ are shared and time-specific spatial log-relative risks, respectively, at location s_{ijt}, a point in area i at location j at time t, $t = \{0, 1\}$. We generate $\xi(\cdot)$ and $\xi_t(\cdot)$ as realizations from a smooth Gaussian process with a Matern($\nu = 0.3, \kappa = 2$) correlation structure, where ν is a range parameter and κ is a smoothness parameter (Figure 4.6). We generate the shared surface between time points, $\xi(\cdot)$, to induce spatio-temporal correlation in the data. We generate an area-level covariate x_{it} from a uniform distribution, and are interested in the parameter β_{xt}, which represents the change in the effect of the covariate across time. The true value for this covariate in our study is $\beta_{xt} = -0.5$.

At each time point, we generate a realization from the continuous Poisson process with rate λ_{ijt}, and aggregate the cases over each area to obtain area-level case counts. We run 2,000 simulations for each scenario described. We model the area-level expected count μ_{it} using Model 3:

$$\log(\mu_{it}) = \log(E_{it}) + \beta_x x_{it} + \beta_t t + \beta_{xt} x_{it} t + \sum_j \sum_l Z_l(s_{ij})(u_l + u_{lt}),$$

where $E_{it} = N_{it}p$; N_{it} is the population size in area i at time t; and the penalized spline terms are identical to those defined in Sections 4.3.2 and 4.4. We simulate our data assuming no residual overdispersion (unless otherwise stated). However, we fit the 'quasi-Poisson' model in our simulations and estimate an overdispersion parameter ϕ, in order to investigate whether identifiability problems arise between spatial variance and overdispersion parameters.

4.5.2 Results of simulation study

First, we examine the power and type I error associated with the test $H_0 : \beta_{xt} = 0$ for Models 1, 2, and 3 for two different settings. In Setting 1, we simulate data from a model with $\sigma = 0.3, \sigma_1 = \sigma_2 = 0.2$ and compare the fits of Models 1, 2, and 3; in Setting 2, we eliminate time-specific heterogeneity by simulating from a model with $\sigma = 0.3, \sigma_1 = \sigma_2 = 0$ and compare the fits of Models 1 and 2.

Figures 4.2 and 4.3 display power curves for Settings 1 and 2, respectively, for testing $H_0 : \beta_{xt} = 0$, when the data contains 64 and 256 areas. These figures illustrate that incorporating temporal correlation in the spatial random effects increases the power to detect differences in a covariate effect across time. The amount of power gained increases as the amount of spatial heterogeneity or temporal correlation in the spatial random effects increases (results not shown) and as the number of areas at each time point decreases.

Misspecifying the model by ignoring this temporal correlation can result incorrect inferences about the parameter β_{xt}, since we are examining the change in a covariate effect across time. In panel (a) in Figures 4.2 and 4.3, we see that the type I error deviates from 0.05 when the spatio-temporal correlation structure is misspecified. When temporal correlation is ignored (Model 2), the type I error is less than 0.05, and the test is overly conservative. In Setting 2, when time-specific heterogeneity is ignored, the type I error is greater than 0.05.

In our simulations, power gains and differences in type I error between the three

models were negligible in the scenario with 1,024 areas (results not shown). Existing models that handle temporal boundary misalignment will perform as well as our proposed model (in terms of the efficiency of β_{xt}). However, these existing approaches are fully Bayesian, and the computational efficiency of our frequentist parameter estimation framework is beneficial when the number of areas is large. Model fitting time can change from days (for alternative Bayesian models) to minutes (using the PQL approximation for parameter estimation).

Tables 4.1, 4.2, and 4.3 show results of the simulation study assessing the sensitivity of the model to: (1) the sparseness of the data, (2) the choice of the range parameter ρ, and (3) the choice of the number of knots G.

In Table 4.1, we assess the performance of the PQL approximation when data are sparse, varying the expected area counts within an area. When testing $H_0 : \beta_{xt} = 0$, the type I error is near 0.05 and 95% Wald CI coverage is near 0.95 regardless of the expected area counts; variance components for the spatial random effects are also unbiased. We conclude that the PQL approximation performs well.

In Table 4.2, we examine sensitivity of the model to the choice of the range parameter ρ (assuming an exponential correlation structure and $\rho_t = \rho$). Zhang (2004) showed that ρ and σ^2 are not jointly identifiable in a spatial GLMM, and so misspecification of ρ leads to inconsistent estimates of σ, σ_1, and σ_2. Point estimates and standard errors of the fixed effects remain accurate, and type I error is near 0.05 when we misspecify the range parameter, insofar as the choice of the range parameter is 'reasonable' (*i.e.* the average 'radius' of an area $< 3/\rho <$ the maximum distance between areas).

In Table 4.4, we evaluate the performance of each model for overdispersion. When modeling non-spatial residual overdispersion, the quasi-Poisson model performs well in the settings with 256 and 1024 areas. With 64 areas, the quasi-Poisson model does not perform as well as the negative-binomial and random-intercept models. In this setting, we could improve somewhat upon the simulation results by using the random-

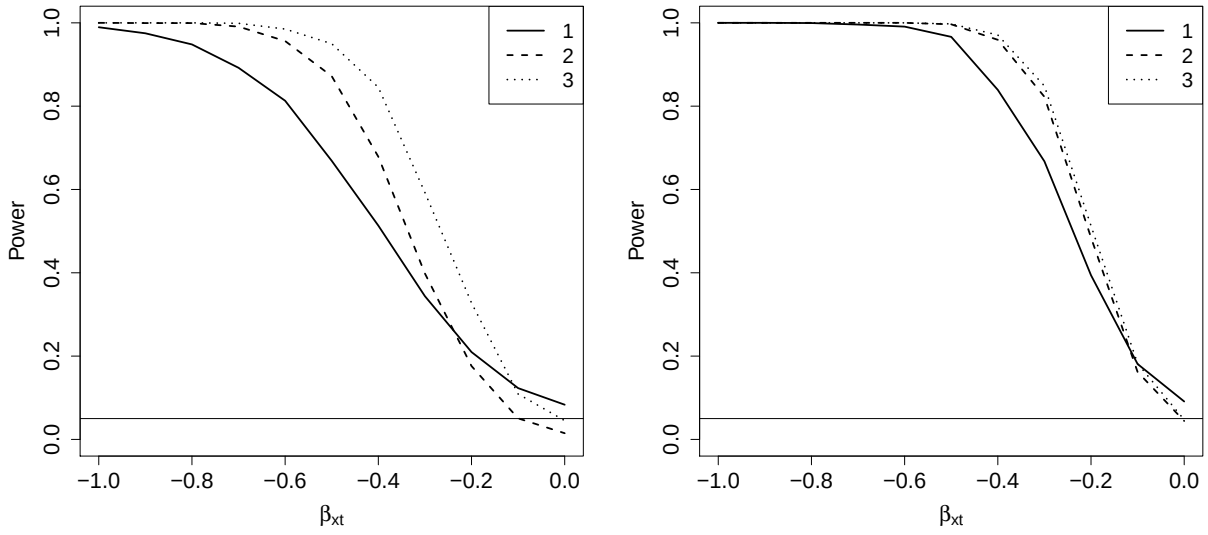


Figure 4.2: Plot of the power for the test $H_0 : \beta_{xt} = 0, H_a : \beta_{xt} \neq 0$, at the $\alpha = 0.05$ level as a function of β_{xt}, when $\beta_x = 1$. (a) 64 and (b) 256 areas. $\sigma = 0.3, \sigma_1 = \sigma_2 = 0.2$. The lines are labeled according to the model that is being fit (*e.g.* '1' corresponds to Model 1.)

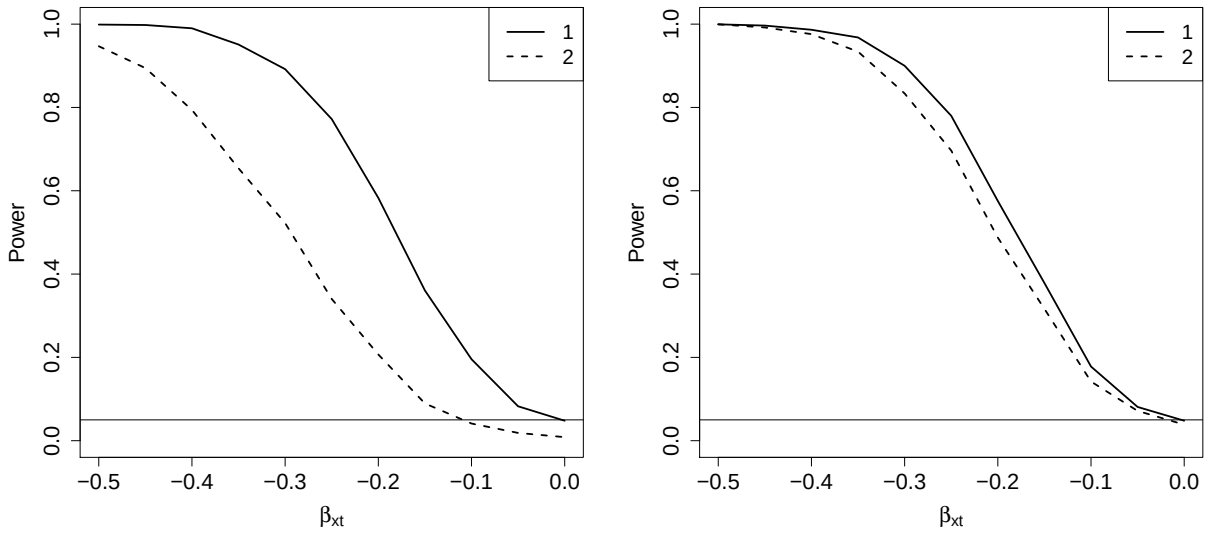


Figure 4.3: Plot of the power for the test $H_0 : \beta_{xt} = 0, H_a : \beta_{xt} \neq 0$, at the $\alpha = 0.05$ level as a function of β_{xt}, when $\beta_x = 1$. (a) 64 and (b) 256 areas. $\sigma = 0.3, \sigma_1 = \sigma_2 = 0$. The lines are labeled according to the model that is being fit (*e.g.* '1' corresponds to Model 1.)

Table 4.1: Determining sensitivity of the model to the sparseness of the data (quantified by the expected count E in each region). We omit results from the $E = 0.1, E = 0.5$ settings from the 64 areas case and $E = 0.1$ from the 256 areas case, because data is sparse enough that some simulations had only 1 or 2 cases, resulting in unstable model estimates. $se = E\{se(\hat{\beta}_{xt})\}$, $sd = SD(\hat{\beta}_{xt})$.

Areas	E	$E(\hat{\beta}_{xt})$	se	sd	95% Cov.	Type I	$E(\hat{\sigma})$	$E(\hat{\sigma_1})/E(\hat{\sigma_2})$	$E(\hat{\phi})$
64	1	-0.517	1.146	1.200	0.934	0.059	0.277	0.266/0.262	0.903
	2	-0.513	0.817	0.851	0.941	0.064	0.300	0.231/0.231	0.915
	5	-0.488	0.533	0.538	0.947	0.062	0.319	0.207/0.211	0.938
	10	-0.503	0.388	0.402	0.939	0.070	0.332	0.215/0.211	0.942
256	0.5	-0.497	0.796	0.802	0.951	0.061	0.295	0.226/0.215	0.935
	1	-0.511	0.568	0.600	0.935	0.054	0.310	0.209/0.202	0.950
	2	-0.501	0.405	0.419	0.944	0.064	0.313	0.205/0.209	0.963
	5	-0.488	0.261	0.268	0.943	0.053	0.307	0.210/0.209	0.979
	10	-0.506	0.190	0.190	0.948	0.061	0.299	0.211/0.209	0.991
1024	0.1	-0.516	0.879	0.916	0.939	0.063	0.289	0.220/0.232	0.935
	0.5	-0.493	0.401	0.404	0.949	0.053	0.309	0.205/0.201	0.969
	1	-0.501	0.285	0.290	0.948	0.050	0.306	0.206/0.211	0.978
	2	-0.510	0.203	0.202	0.950	0.054	0.299	0.210/0.207	0.986
	5	-0.499	0.130	0.129	0.951	0.052	0.282	0.207/0.204	1.002
	10	-0.502	0.093	0.093	0.955	0.045	0.273	0.203/0.199	1.017

Table 4.2: Determining sensitivity of the model to choice of the range parameter. $se = E\{se(\hat{\beta}_{xt})\}$, $sd = SD(\hat{\beta}_{xt})$.

Areas	ρ	$E(\hat{\beta}_{xt})$	se	sd	95% Cov.	Type I	$E(\hat{\sigma})$	$E(\hat{\sigma_1})/E(\hat{\sigma_2})$	$E(\hat{\phi})$
	3	-0.509	0.136	0.142	0.938	0.058	0.440	0.266/0.270	1.171
	5	-0.507	0.131	0.134	0.942	0.056	0.326	0.218/0.219	1.000
64	10	-0.507	0.137	0.139	0.942	0.060	0.324	0.231/0.229	0.888
	20	-0.508	0.157	0.158	0.936	0.068	0.481	0.333/0.331	0.937
	40	-0.510	0.197	0.199	0.942	0.053	1.172	0.622/0.652	1.995
	3	-0.500	0.100	0.099	0.956	0.058	0.324	0.227/0.226	1.076
	5	-0.500	0.101	0.099	0.954	0.055	0.273	0.200/0.198	1.075
256	10	-0.500	0.105	0.101	0.957	0.051	0.261	0.199/0.196	1.124
	20	-0.501	0.117	0.107	0.968	0.041	0.345	0.253/0.250	1.369
	40	-0.502	0.141	0.120	0.980	0.027	0.686	0.402/0.412	2.091
	3	-0.498	0.093	0.093	0.949	0.049	0.325	0.228/0.227	1.019
	5	-0.498	0.093	0.093	0.949	0.049	0.274	0.200/0.198	1.019
1024	10	-0.498	0.094	0.093	0.951	0.049	0.260	0.200/0.195	1.031
	20	-0.498	0.097	0.094	0.956	0.045	0.344	0.255/0.251	1.082
	40	-0.499	0.105	0.098	0.962	0.038	0.669	0.422/0.423	1.263

Table 4.3: Determining sensitivity of the model to choice of the number of knots parameter. $se = E\{se(\hat{\beta}_{xt})\}$, $sd = SD(\hat{\beta}_{xt})$.

Areas	G	$E(\hat{\beta}_{xt})$	se	sd	95% Cov.	Type I	$E(\hat{\sigma})$	$E(\hat{\sigma_1})/E(\hat{\sigma_2})$	$E(\hat{\phi})$
	16	-0.497	0.221	0.158	0.993	0.010	0.417	0.195/0.204	5.308
	25	-0.497	0.195	0.148	0.989	0.014	0.378	0.188/0.196	3.819
64	36	-0.497	0.174	0.139	0.981	0.016	0.346	0.189/0.193	2.777
	49	-0.495	0.144	0.127	0.970	0.034	0.396	0.209/0.215	1.681
	64	-0.496	0.131	0.127	0.952	0.050	0.319	0.215/0.213	0.980
	36	-0.502	0.113	0.105	0.967	0.027	0.352	0.220/0.217	1.455
	64	-0.502	0.105	0.102	0.956	0.045	0.303	0.208/0.202	1.198
256	100	-0.501	0.101	0.101	0.953	0.046	0.273	0.200/0.194	1.073
	144	-0.501	0.101	0.101	0.953	0.046	0.273	0.200/0.194	1.073
	196	-0.501	0.101	0.101	0.953	0.046	0.273	0.200/0.194	1.073
	25	-0.504	0.099	0.093	0.952	0.048	0.391	0.241/0.245	1.184
	49	-0.505	0.095	0.092	0.953	0.054	0.330	0.221/0.219	1.079
1024	100	-0.505	0.093	0.091	0.950	0.060	0.274	0.201/0.199	1.018
	144	-0.505	0.092	0.091	0.948	0.062	0.254	0.194/0.191	0.999
	225	-0.507	0.106	0.106	0.943	0.051	0.260	0.214/0.126	0.980

Table 4.4: Comparing models for residual overdispersion, where ϕ denotes an overdispersion parameter (with different meaning for each of the different models). QP, NB, and RI denote the quasi-Poisson model, the negative-binomial model, the random intercept model, respectively. $\hat{\phi}$ represents the traditional residual overdispersion parameter in the QP model; the scale parameter for the negative binomial model for count data in the NB model; and the variance of the random intercepts in the RI model. We did not consider the random intercept model for the 1024 area case, because the other models performed sufficiently well and were substantially faster to fit.

Areas	Model	$E(\hat{\beta}_{xt})$	$E\{se(\hat{\beta}_{xt})\}$	$SD(\hat{\beta}_{xt})$	95% Cov.	Type I	$E(\hat{\phi})$	% Converge
64	NB	-0.499	0.466	0.484	0.940	0.058	0.034	0.812
	QP	-0.490	0.459	0.489	0.926	0.069	1.362	1.000
	RI	-0.497	0.466	0.470	0.947	0.060	0.035	0.852
256	NB	-0.493	0.223	0.235	0.936	0.063	0.037	0.222
	QP	-0.501	0.226	0.233	0.944	0.061	1.456	1.000
	RI	-0.496	0.228	0.231	0.944	0.060	0.038	0.910
1024	NB	-0.489	0.111	0.112	0.953	0.064	0.042	0.268
	QP	-0.502	0.112	0.114	0.949	0.052	1.504	1.000

intercept model (which is the 'correct model' in our simulation study). Convergence issues arose with the negative binomial model in our simulations, whereas the random-intercept model converges $> 85\%$ of the time. The quasi-Poisson model converges $> 99\%$ of the time in simulation.

Additionally, we note that the variance parameters σ^2, σ_1^2, and σ_2^2 are relatively unbiased when the range parameter is correctly specified (Table 4.2), suggesting that the model which allows for a common surface and time-specific spatial surfaces is identifiable with sufficient data. This result is consistent with Wager et al. (2004) and Coull et al. (2001), who fit similar models to Model 3.

Further, we find that when data are generated from a smooth underlying surface, a model with as few as ~ 64 knots will perform as well as models with higher knot choices (Table 4.3). When the underlying spatial surface is less smooth, more knots are required to appropriately model the surface. For instance, if the range of the spatial correlation is less

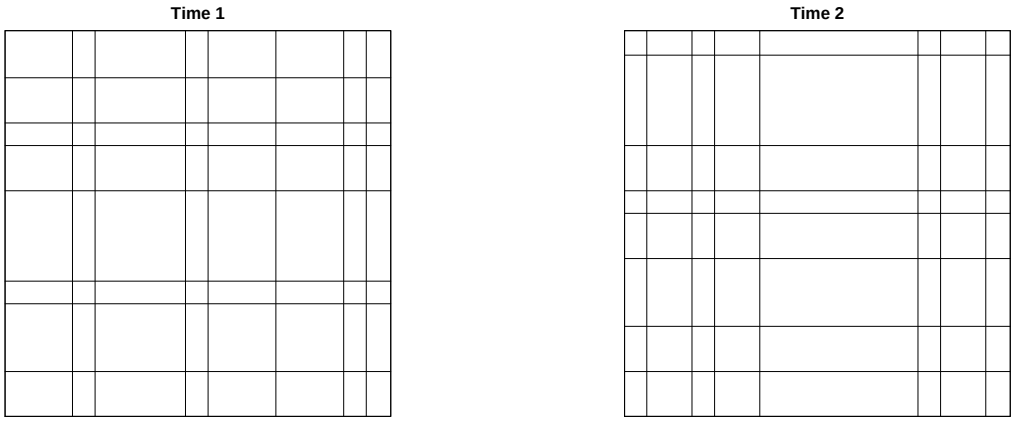


Figure 4.4: Non-regular, misaligned grid used in the simulation study with 64 areas.

than the minimum distance between knots, oversmoothing of the spatial surface occurs. When we do not include enough knots in the model (for instance, the scenario with 64 areas and $G < 64$), the data appears highly overdispersed ($\hat{\phi} >> 1$) and the standard error of $\hat{\beta}_{xt}$ is underestimated.

For the scenarios with 64 and 256 areas, we repeat our simulations assessing sensitivity of the model to knot selection, choice of range parameter, and the number of design points using a non-regular, misaligned grid, shown in Figure 4.4. We exclude the 1,024 case because we observed the greatest power differences and sensitivity to parameter choices for the 64 and 256 area scenarios; additionally, the setting with 1,024 areas better approximates a regular grid. When estimating fixed effects and variance parameters, the model was not sensitive to the choice of the number of design points (results not shown).

In Table 4.5, we examine how well the model predicts the area-specific relative risks as a function of the number of knots included in the model, when the data is no longer on a regular grid. We present the average mean-squared error (defined in Section 4.8), as well as estimates of β_{xt} and $se(\hat{\beta}_{xt})$. The results from the irregular grid are nearly identical

Table 4.5: Comparing the performance of our model when data is simulated on a regular versus an irregular, misaligned grid. We compare the models for different choices of the number of knots, for the scenarios with 64 and 256 areas. $se = E\{se(\hat{\beta}_{xt})\}$, $sd = SD(\hat{\beta}_{xt})$.

		Misaligned Irregular Grid				Regular Grid			
Areas	G	$E(\hat{\beta}_{xt})$	se	sd	MSE	$E(\hat{\beta}_{xt})$	se	sd	MSE
64	36	-0.501	0.173	0.164	0.013	-0.499	0.171	0.150	0.012
	49	-0.500	0.161	0.155	0.009	-0.498	0.139	0.136	0.006
	64	-0.499	0.155	0.153	0.008	-0.497	0.125	0.135	0.005
256	100	-0.502	0.099	0.098	0.008	-0.502	0.101	0.100	0.009
	144	-0.502	0.098	0.098	0.007	-0.502	0.099	0.100	0.008
	196	-0.502	0.098	0.097	0.007	-0.502	0.097	0.100	0.008

to the regular grid, with a small inflation in the MSE when the data contains only 64 areas. The results from our simulations suggest that the model results and model validity do not change substantially, regardless of whether the data is misaligned over time.

Using data simulated on the irregular, misaligned grid, we evaluate how well our model performs when the range parameter changes across time, but we fit a model assuming that the range is constant over time (see Section 4.8 for detailed description of data generation). Under these conditions, we still obtain valid estimates of β_{xt} and $se(\hat{\beta}_{xt})$ (Table 4.6). Once again, correct specification of the range parameter is not important in obtaining valid model results, due to the lack of identifiability between the spatial variance parameters and the range parameter.

Lastly, the estimated overdispersion parameter $\hat{\phi}$ in our simulations is often less than 1, suggesting that data are underdispersed. Specifically, data appear underdispersed when choice of G (number of knots) is high as well as when data are sparse (Tables 4.1 and 4.3). In this situation, unless there is a plausible reason for why the data are underdispersed, we recommend fixing $\phi = 1$ or adjusting the number of knots such that $\hat{\phi} \approx 1$, as suggested in Wager et al. (2004). Based on our simulations, when estimates of ϕ are less than 1, the model performance improves when we fix $\phi = 1$ and do not estimate an

Table 4.6: Comparing the performance of our model when the range parameter changes across time and when the range parameter is fixed over time, fitting a model which assumes the latter is true. Data is simulated on an irregular, misaligned grid, for the scenarios with 64 and 256 areas. $se = E\{se(\hat{\beta}_{xt})\}$, $sd = SD(\hat{\beta}_{xt})$.

		Range parameter changes				Fixed range parameter			
Areas	ρ	$E(\hat{\beta}_{xt})$	se	sd	MSE	$E(\hat{\beta}_{xt})$	se	sd	MSE
	3	-0.493	0.165	0.178	0.014	-0.498	0.151	0.149	0.009
64	5	-0.494	0.166	0.179	0.010	-0.497	0.155	0.153	0.008
	10	-0.495	0.182	0.192	0.012	-0.496	0.175	0.173	0.010
	3	-0.501	0.104	0.102	0.012	-0.503	0.098	0.097	0.007
256	5	-0.502	0.105	0.102	0.012	-0.503	0.099	0.097	0.008
	10	-0.502	0.107	0.103	0.012	-0.503	0.101	0.099	0.009

overdispersion parameter (results not shown).

4.6 Analysis of the Los Angeles Breast Cancer Data

Using the spatio-temporal model described in Section 4.4, we re-analyze the Los Angeles cancer data presented in Krieger et al. (2006), restricting our attention to the time periods 1988-1992 and 1998-2002. Descriptive statistics for the LA cancer data are shown in Table 5.1. The number of census tracts in LA county was 1,642 in 1990 and 2,056 in 2000; the total population count of women over 15 years old was 3,492,249 in 1990 and 3,625,360 in 2000. Following standard practice for cancer incidence rates centered around a census (Boyle and Parkin, 1991), we estimate person-time by assuming that the population counts are constant within each 5-year time period (1988-1992 and 1998-2002) and multiply the decennial population counts from the censuses by 5.

Age at diagnosis is categorized into 8 groups: 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, and $84+$. Data are available for white non-Hispanic, Hispanic, black non-Hispanic, and Asian Pacific Islander populations. We use internal standardization to calculate the expected number of breast cancer cases by CT for each time period. We

analyze the data combining all race/ethnicities (standardizing by age and race/ethnicity) and for each race/ethnicity group individually (standardizing only by age). Chen et al. (2008) emphasize that it may not be appropriate to assume a common spatial effect across racial/ethnic groups due to patterns of racial/ethnic segregation. We report results for all race/ethnicities combined and for the two largest subgroups, White non-Hispanics and Hispanics.

We select the percent of the population below the poverty level in a CT as our area based socioeconomic measure (ABSM). Following Krieger et al. (2006), we model the relationship between the ABSM and the log-relative risks associated with pre-determined epidemiologically meaningful poverty groups. Therefore, we model the percent of the population below poverty as a 5-level categorical variable as follows: (a) among census tracts with $< 5\%$ poverty, we distinguish between those with $\geq 10\%$ high income households (8.3% of CTs) and $< 10\%$ high income households (9.2% of CTs); and (b) among the remaining census tracts, we distinguish between those with $5.0 - 9.9\%$ (23.6% of CTs), $10.0 - 19.9\%$ (26.6% of CTs), and $\geq 20\%$ poverty (the federal definition of a "poverty area" and 32.4% of CTs). High income households are defined as ≥ 4 times the US median household income.

To model spatial variability, we define $\rho = 15/\Delta$ based on epidemiological plausibility, where Δ is the maximum distance between CTs in LA county. We use 30 design points per CT and select 100 knots throughout the study region using the space filling design described in Johnson et al. (1990) and implemented in the R package FIELDS.

Let Y_{it} denote the observed number of incident breast cancer cases in CT i at time t,

Table 4.7: Descriptive Statistics for LA Breast Cancer Data. Median (IQR) are presented.

	Population Size	Observed Cases	Expected Cases
All	9150 (6980, 12005)	12 (7, 18)	12.2 (8.2, 17.4)
White non-Hispanic	3060 (710, 6061.3)	6 (1, 13)	7.0 (1.8, 13.6)
Hispanic	2742.5 (1120, 5180)	2 (1, 3)	1.7 (0.8, 3.1)

and assume $Y_{it} \sim Poisson(\mu_{it})$. We fit the model (analogous to Model 3):

$$\log(\mu_{it}) = \log(E_{it}) + \beta_0 + \boldsymbol{\beta^p pov}_{it} + \beta^t I_{t=2000} + \boldsymbol{\beta^{pt} pov}_{it} I_{t=2000} + \sum_j \sum_l Z_l(s_{ij})(u_l + u_{lt})$$

where $\boldsymbol{pov}_{it}$ is a 4×1 indicator variable for the poverty category of area i at time t; and $\sum_j \sum_l Z_l(s_{ij})(u_l + u_{lt})$ is the spatio-temporal spline term, defined in Model 3 in Section 4.4. For model identifiability, we use the '$> 20\%$ poverty' category as the reference category.

For each race/ethnicity, fitting Model 3 takes approximately 15 minutes using the glmmPQL function in R and 3 minutes using PROC GLIMMIX in SAS. Based on the results from Model 3 in Table 4.8, the socioeconomic gradient in breast cancer does not appear to be decreasing over the time period studied. Instead, consistent with the findings in Krieger et al. (2006), we observed that the IRR remained stable over time in the different racial/ethnic groups, and that the socioeconomic gradient was smaller among the white non-Hispanic women (among whom the "catch up" may have already occurred), and greater among Hispanic women, for whom cancer risk factors may still exhibit strong socioeconomic patterning.

In Table 4.9, we examine the estimated spatial variance parameters and compare the results from Models 2 and 3. The standard errors of the estimated fixed effects $\hat{\boldsymbol{\beta}}^{\boldsymbol{pt}}$ from Model 3 are consistently smaller than those in Model 2 in all analyses. In our analysis, incorporating correlation between the spatial random effects across time results in a substantial increase in power. Specifically, in the combined racial/ethnic group analysis, we have stronger evidence that there exists a change in the socioeconomic gradient of breast cancer over time using Model 3 ($p = 0.03$) versus Model 2 ($p = 0.15$). In Figure 4.5, we plot the common residual spatial surface across both time points for all races combined, as well as the residual spatial surface from 1990, estimated using Model 3 (note that we do not detect any residual spatial variability for the 2000 time point, as $\hat{\sigma_2} \approx 0$). The similarity between the spatial surfaces across time drives the gain in efficiency obtained when we use Model 3.

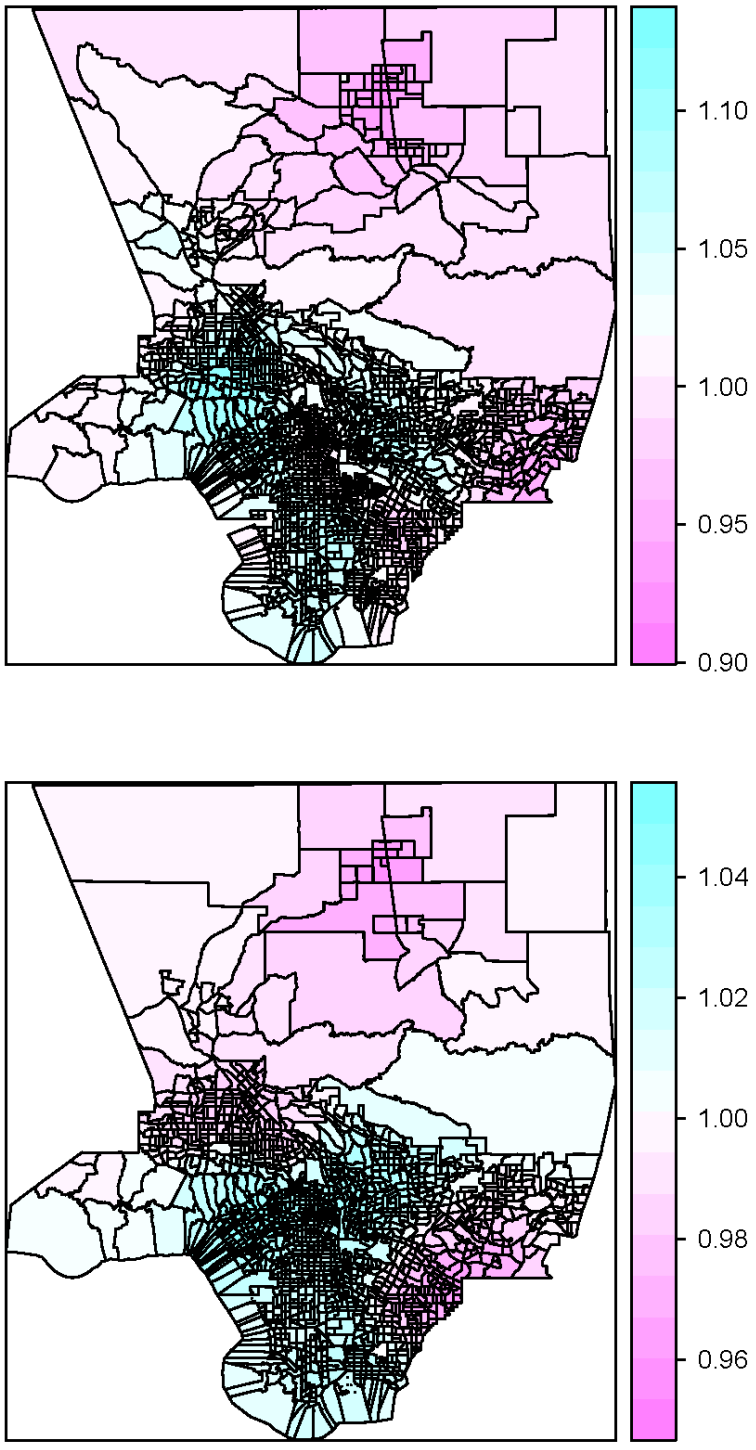


Figure 4.5: Plot of the common spatial residual relative risk surface for the 1990 and 2000 time periods in LA county; and the additional time-specific spatial residual relative risk surface for the year 1990 in LA County.

Table 4.8: Results from Los Angeles cancer data analysis. Incidence rate ratios relative to the $> 20\%$ poverty category are shown for each time period and race/ethnicity group, with Wald p-values testing whether the log-IRR changes across time for each poverty category and for all categories combined.

	Category	IRR 1988-1992	IRR 1998-2002	p-value
All	$> 20\%$	1	1	-
	$10 - 20\%$	1.09 (1.05, 1.14)	1.09 (1.05, 1.13)	0.9829
	$5 - 10\%$	1.13 (1.08, 1.18)	1.18 (1.14, 1.22)	0.1592
	$< 5\%$ & $< 10\%$ high inc.	1.15 (1.09, 1.20)	1.28 (1.22, 1.34)	0.0056
	$< 5\%$ & $\geq 10\%$ high inc.	1.25 (1.19, 1.30)	1.28 (1.23, 1.33)	0.5053
	Wald test, 4df			0.0281
White	$> 20\%$	1	1	-
	$10 - 20\%$	1.01 (0.93, 1.08)	0.992 (0.93, 1.06)	0.755
	$5 - 10\%$	1.03 (0.96, 1.11)	1.073 (1.01, 1.14)	0.439
	$< 5\%$ & $< 10\%$ high inc.	1.07 (0.98, 1.15)	1.150 (1.07, 1.23)	0.178
	$< 5\%$ & $\geq 10\%$ high inc.	1.15 (1.07, 1.23)	1.190 (1.12, 1.26)	0.501
	Wald test, 4df			0.2654
Hispanic	$> 20\%$	1	1	-
	$10 - 20\%$	1.16 (1.07, 1.25)	1.23 (1.16, 1.31)	0.2752
	$5 - 10\%$	1.33 (1.22, 1.44)	1.43 (1.34, 1.52)	0.2770
	$< 5\%$ & $< 10\%$ high inc.	1.38 (1.23, 1.53)	1.79 (1.64, 1.94)	0.0119
	$< 5\%$ & $\geq 10\%$ high inc.	1.63 (1.43, 1.83)	1.62 (1.43, 1.81)	0.9714
	Wald test, 4df			0.1419

Table 4.9: Results from LA breast cancer data analysis, comparing spatial variance parameters and p-values testing $H_0 : \boldsymbol{\beta^{pt}} = 0$ between Model 2 (spatial random effects are independent across time) and Model 3 (allows for temporal dependence in spatial random effects).

Race/Ethnicity	Model	$\hat{\sigma}$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\sqrt{\hat{\phi}}$	p-value
All	3	0.0870	0.0461	0.000	1.109	0.0281
	2	-	0.1008	0.0746	1.110	0.1497
White	3	0.1304	0.000	0.000	1.162	0.2654
	2	-	0.1572	0.0971	1.155	0.2904
Hispanic	3	0.1290	0.000	0.000	1.061	0.1419
	2	-	0.1411	0.0905	1.061	0.1906

4.7 Discussion

Motivated by the temporal boundary misalignment issues in the Los Angeles breast cancer incidence study, we develop an area-level disease mapping model that incorporates spatio-temporal correlation in the presence of temporal boundary misalignment. Anyone using U.S. census data from more than one decade inevitably encounters the same temporal boundary misalignment issues that we face. Previous solutions to this problem are computationally intensive and ignore temporal correlation in the spatial random effects, potentially leading to inefficient inferences.

The proposed model does require selecting a parametric form for the correlation structure for the underlying continuous relative risk surface. Our simulation study suggests that the exponential correlation structure performs well and that the choice of the range parameter ρ is not too important. While fixing the range parameter may seem arbitrary, Zhang (2004) prove that, in spatial GLMMs, it is impossible to consistently estimate ρ and the variance parameter σ^2, but that the ratio σ^2/ρ is both more stable and more important to interpolation than the individual parameters. Therefore, fixing one parameter (ρ) and estimating the other (σ^2) should provide a consistent estimate of the spatial random effects.

While we have emphasized the usefulness of our model in addressing the temporal boundary misalignment problem, it is important to note that this new method will be a very useful and computationally efficient alternative to the popular fully-Bayesian disease mapping models for data collected at a single time point. Most disease mapping applications in the literature use study regions containing only a few hundred areas, and fitting fully Bayesian models is feasible in such cases. For larger datasets with thousands of areas, which are becoming more common in epidemiological applications, these Bayesian models are more difficult to implement. By using a PQL approximation to maximum likelihood inference, we reduce the computation time from hours to minutes for our dataset and avoid any issues associated with model convergence and prior selection. Our method is also easy to program in standard software (SAS and R), filling a gap in the available software for fitting GLMMs with area-level spatial correlation.

Furthermore, while constructed in a different manner, the area-level spatial prior in our model has the same interpretation as that proposed in Kelsall and Wakefield (2002). Their model is often cited in disease mapping reviews as a good option for modeling area-level spatial correlation, as it seems appropriate to model the area-level relative risk as arising from a continuous underlying surface. However, we could not find any articles that use this method in practice, presumably due to the challenges associated with model fitting. We hope that the simplicity of our model will facilitate its use in practice.

In the present model for spatio-temporal variability, following Wager et al. (2004), we assume that the correlation between the log-relative risk at a given location at different time points is the same. When boundaries are aligned across time, this corresponds to the assumption that $\mathrm{Corr}\{\log(\mu_{it_j}, \mu_{it_k})\} = c$, where c is a constant, for all time points j, k. When data are available at only two time points, this model is appropriate. When data are available at more than two time points, one might develop more sophisticated longitudinal extensions of this model that induce more correlation between occasions closer together in time. One viable option is using a model-based approach and incorporating spatio-temporal correlation through placing relevant priors on the random effects $\{u_{lt}\}$,

such as $u_{lt} \sim N\{0, \Sigma(\rho_t)\}$. For instance, we could specify an $\mathrm{AR}(1)$ prior on $\{u_{lt}\}$. Fitting a frequentist version of this model in standard software is also of interest.

Using the Los Angeles County breast cancer incidence data, we find no clear evidence supporting the hypothesis that the socioeconomic gradient in breast cancer incidence is decreasing over time, consistent with the findings in Krieger et al. (2006). Results were robust to the choice of model parameters, including as the range parameter, number of knots, or the ABSM included in the regression model.

4.8 Detailed description of the simulation study

To construct our datasets, we generate a continuous log-relative risk surface and a continuous log-population density surface using a Gaussian random field on a 512×512 pixel grid. The common spatial variability is induced by spatial log-relative risk surfaces $\xi(s)$, generated from $\boldsymbol{\xi} \sim \mathrm{GRF}\{\mathbf{0}, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\nu})\}$; similarly, the time-specific log-risk surfaces are generated from $\boldsymbol{\xi_t} \sim \mathrm{GRF}\{\mathbf{0}, \sigma_t^2\boldsymbol{\Sigma}(\boldsymbol{\nu})\}$. We generate relatively smooth surfaces, choosing the correlation structure $\boldsymbol{\Sigma_t}(\boldsymbol{\nu}) = \mathrm{Matern}(0.3, 2)$ on a $(0,1) \times (0,1)$ grid (Figure 4.6). The population density surface is generated similarly, using a Matern(0.3,2) log-population density surface, but standardized to have a total population of ~ 9.5 million people over the entire grid (similar to the data application).

We divide the grid into 64, 256, or 1024 square blocks (areas) and simulate data on this grid at two time points. In order to reflect the attributes of the motivating L.A. breast cancer dataset, the incidence p is 0.11 cases per 100 person-years. In the scenario with 1024 areas, the average number of cases and total population per area are 10 and 9,000, respectively; these values increase as the number of areas decreases, in order to maintain the same p throughout the analysis and to illustrate how the results change when level of aggregation of the data changes. We do not use age-specific disease rates in our simulation study; the expected number of cases in an area is crudely defined as the product of the disease incidence and the population size within an area.

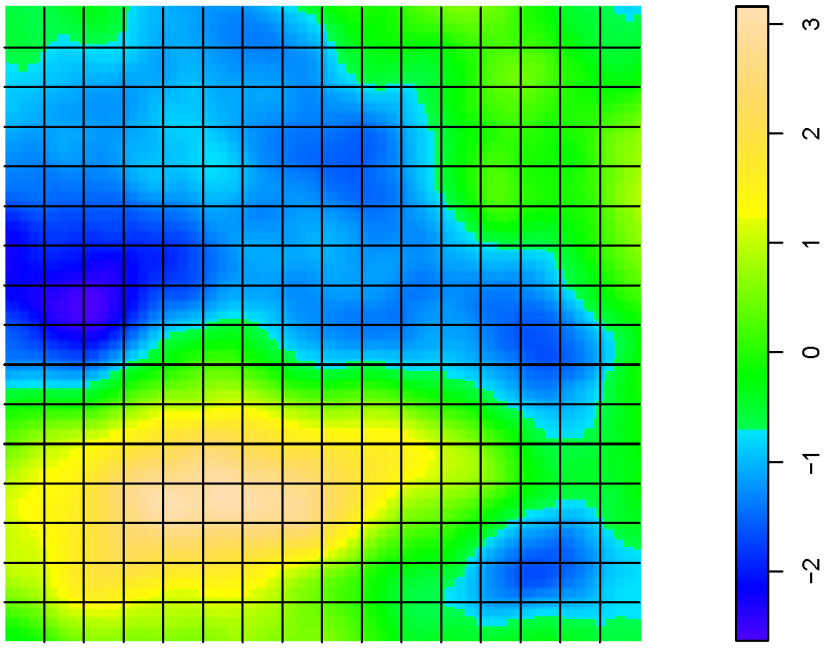


Figure 4.6: Realizations from Gaussian process with Matern(0.3,2) correlation structure on $(0,1) \times (0,1)$ grid, divided into 256 areas.

We simulate a Poisson process with intensity λ_{ijt} at each point location on the grid, where the log-intensity is (analogous to Model 3 in Section 4.4):

$$\log(\lambda_{ijt}) = \log(E_{ijt}) + \beta_x x_{it} + \beta_t t + \beta_{xt} x_{it} t + \xi(s_{ijt}) + \xi_t(s_{ijt}),$$

where s_{ijt} is a point in area i at location j at time t, $t = \{0, 1\}$, and E_{ijt} is the expected number of cases at time t in area i at point j.

Unless stated otherwise, the true parameter values in the simulation study are: $\beta_x = 1, \beta_t = 0.2, \beta_{xt} = -0.5, \sigma = 0.3, \sigma_1 = \sigma_2 = 0.2$. To assess type I error, we fix $\beta_{xt} = 0$. The area-level covariates x_{it} are generated from a $\text{Unif}(0, 0.5)$ distribution, to reflect a poverty indicator such as percent of the population below poverty. Spatial random effects s_{ij} and

s_{ijt} are generated from Gaussian processes as described above, such that the point-wise relative risk attributed to underlying spatial heterogeneity lies between 0.71 and 1.40 50% of the time and between 0.37 and 2.66 95% of the time.

This data generating mechanism does not induce any residual overdispersion into the data. To assess the appropriate model for residual overdispersion, we induce overdispersion in our data by introducing a covariate $x_{O,it} \sim N(0, 0.2^2)$ into the data generating mechanism and omitting this covariate when fitting the model.

To obtain area-level population counts from the underlying population density data, we integrate over the density surface. Using the population density surface and log-relative risk from the above model, we generate an area-level case counts from a realization of the underlying Poisson process model with rate λ_{ijt} at location i in area j at time t. We run 2,000 simulations for each scenario described.

We then model the area-level expected count μ_{it} using Model 3:

$$\log(\mu_{it}) = \log(E_{it}) + \beta_x x_{it} + \beta_t t + \beta_{xt} x_{it} t + \sum_j \sum_l Z_l(s_{ij})(u_l + u_{lt}),$$

where $E_{it} = N_{it}p$; N_{it} is the population size in area i at time t; and the penalized spline terms are identical to those defined in Sections 4.3.2 and 4.4. We assume the underlying spatial surface follows a Gaussian process with an exponential correlation structure with $\rho = 5$ (note that we are misspecifying the correlation structure throughout the simulation study, as we generated data from a Matern correlation function; this misspecification did not affect our results). For settings with more than 64 areas, we fit a reduced-rank model with 100 knots spaced evenly across the grid to represent spatial heterogeneity. For the 64 area case, we use 64 knots. We account for residual overdispersion using the 'quasi-Poisson' model. Though we simulate our data assuming no residual overdispersion (except for the simulations in Table 4.4, which include residual overdispersion), we include an overdispersion parameter ϕ in all of the models we fit to investigate whether identifiability problems arise between spatial variance parameters and residual overdispersion parameters.

In the Los Angeles breast cancer incidence analysis, we are specifically interested in testing whether a covariate effect changes across time. Using the model above, this corresponds to testing the null hypothesis $H_0 : \beta_{xt} = 0$, and we focus on the parameter β_{xt} throughout the simulation study.

As a supplement to Section 4.5.2, which describes results of the simulation study, we provide Tables 4.1-4.6 displaying simulation results assessing various aspects of our model. Specifically, we assess the performance of the PQL approximation procedure when data are sparse (Table 4.1). We also report the sensitivity of the model to choice of the range parameter (Table 4.2); the choice of the number of knots (Table 4.3); and the choice of the method used to model residual overdispersion (Table 4.4). All simulations are performed with the true value $\beta_{xt} = -0.5$, except to calculate the Type I error, in which case $\beta_{xt} = 0$. For each of the tables below, we present the model-based standard errors for $\hat{\beta}_{xt}$ and the Monte Carlo standard deviations of $\hat{\beta}_{xt}$. Columns 4 and 5 show 95% Wald confidence interval coverage when $\beta_{xt} = -0.5$ and type I error when $\beta_{xt} = 0$, respectively. In the simulation, recall that true values of $\sigma, \sigma_1,$ and σ_2 are 0.3, 0.2, and 0.2. The last column in the tables shows the average estimate of the overdispersion parameter ϕ.

Tables 4.1-4.4 present the results of simulations when we generate data on a regular grid. For the scenarios with 64 and 256 areas, we run additional simulations assessing sensitivity of the model to knot selection, choice of range parameter, and the number of design points using data simulated on a non-regular, misaligned grid (Figure 4.4). Results from the simulations using the non-regular, misaligned grid are in Tables 4.5 and 4.6.

In Tables 4.5 and 4.6, we examine the estimated area-level relative risks for different scenarios, by calculating the average mean-squared error for the area-level relative risk estimates:

$$MSE = E\left\{ \tfrac{1}{Mt} \sum_{i,t} (\hat{R}_{it} - R_{it})^2 \right\}.$$

R_{it} and $\hat{R}_{it}$ are the true and estimated area-level relative risks, respectively, for area i at time t and M is the number of areas.

In Table 4.6, we assess whether the model is sensitive to the misspecification of the range parameter. When generating the data, we fix the range parameter in one set of simulations (using a Matern(0.3, 2) correlation structure on the $(0, 1) \times (0, 1)$ grid) and allow this parameter to change over time in another set (using 3 different correlation structures: Matern(0.3,2) for the shared spatial surface, Matern(0.1,2) and Matern(0.5, 2) surfaces for the time-specific spatial heterogeneity). In Table 4.6, we compare the fixed effects estimates and standard errors, as well as the average MSE, from these simulations. The MSE is lower when we fix the range parameter, but this result is likely an artifact of how we simulate the data. Specifically, when the range parameter changes across time, we generate data using a smaller range parameter at one time point than when the range parameter is fixed; the smaller range parameter results in a more heterogeneous simulated surface that is more difficult to predict, inflating the MSE.

Addressing statistical uncertainty associated with denominator uncertainty and temporal misalignment in disease mapping studies

Lauren Hund and Brent Coull

Department of Biostatistics

Harvard School of Public Health

5.1 Introduction

Temporal disease mapping applications in public health are becoming progressively more complicated, as the size and complexity of available data increases. We aim to map changes in breast cancer incidence over time in Los Angeles County and relate these changes to socioeconomic status. Using data from 1980-2000, there are approximately 35,000 data points. In US census data, shifting census tract boundaries over time cause area-to-area temporal boundary misalignment (Chen et al., 2008; Hund et al., 2012). Additionally, accounting for spatio-temporal correlation disease mapping models is difficult when datasets are large and boundary misalignment occurs.

Another issue that arises frequently in temporal disease mapping applications is missing census tract population counts within age and race/ethnicity strata at the intercensal years (Best and Wakefield, 1999). The issue of how large a role uncertainty in denominators plays in disease mapping studies is an open question. Phipps et al. (2005) illustrate how intercensal population projection errors within age and race/ethnicity strata can induce bias in the estimation of breast cancer incidence rates in counties in California. Smith and Shahidullah (1995) quantify population projection errors for census tracts in Florida using past census data to project census tract counts, comparing their projections to current census data.

Predicting intercensal population counts introduces additional layers of uncertainty into disease mapping models. Best and Wakefield (1999) propose a Bayesian framework for incorporating uncertainty into disease mapping models when intercensal denominators are unknown. The Best and Wakefield (1999) interpolation model for intercensal counts may not be optimal when predicting population counts in different age and race/ethnicity strata; further, the method becomes more computationally intense as the number of census tracts increases and may not be feasible when datasets contain thousands of census tracts at each time point. However, to our knowledge, no other methods exist for incorporating uncertainty in intercensal count projections in disease mapping

models.

In this paper, we discuss common issues in spatio-temporal disease mapping applications, specifically addressing uncertainty in intercensal denominator projections and temporal boundary misalignment. We propose a new framework for predicting intercensal population counts and for assessing the impact of uncertainty in these predictions on health effects analyses. We then quantify the statistical uncertainty associated with population count uncertainty when estimating the relationship between socioeconomic status and breast cancer incidence. In Section 5.2, we introduce the Los Angeles County breast cancer incidence dataset. In Section 5.3, we propose various modeling frameworks for intercensal denominator interpolation, and we evaluate the performance of these models in simulation in Section 5.4. In Section 5.5, we construct a general framework for addressing spatial misalignment in regression models; we apply this model to the LA breast cancer dataset in Section 5.6 to assess changes in the socioeconomic gradient in breast cancer incidence between 1980 and 2000.

5.2 Assessing socioeconomic gradients in breast cancer incidence in LA county

Breast cancer is typically characterized as a disease of affluence, but Krieger et al. (2006) predict that incidence rates may be "catching up" among poorer women in more affluent countries. We investigate the hypothesis that the socioeconomic gradient in breast cancer incidence is decreasing over time by examining data associations between socioeconomic measures and breast cancer incidence rates between 1980 and 2000 in Los Angeles County, CA. For other analyses of and descriptions of this dataset, see Krieger et al. (2006); Chen et al. (2008); and Hund et al. (2012).

We obtained the breast cancer case data from the Los Angeles Cancer Surveillance Program cancer registry. We appended the census tract geocode to each cancer registry record, based on the location and date of residence at diagnosis. We obtained population

Table 5.1: Descriptive Statistics for LA Breast Cancer Data. Median (IQR) are presented.

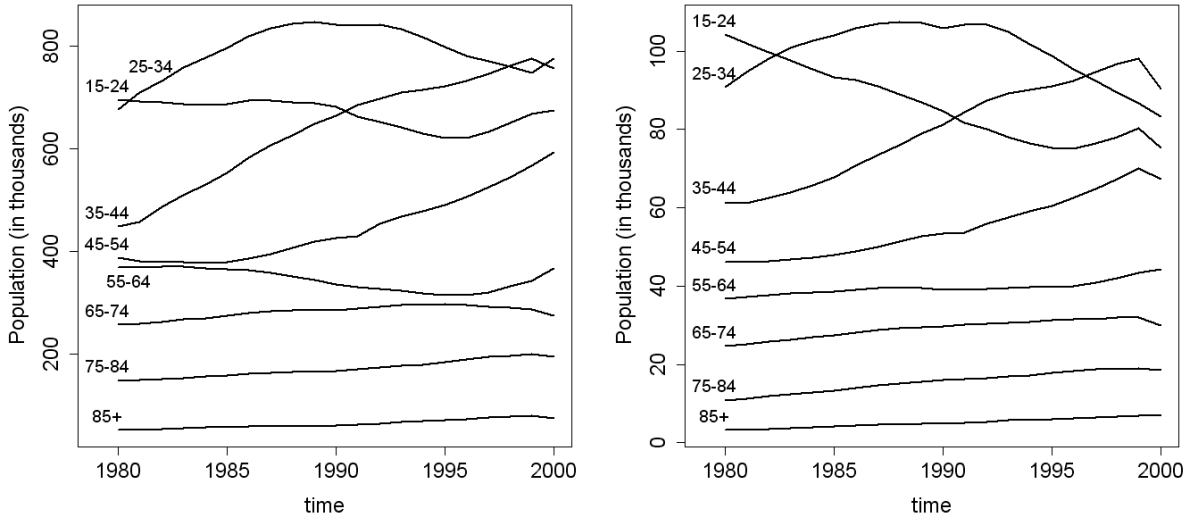
	Population	Expected cases	Observed cases	% below poverty
Total women				
1980	1790 (1288, 2335)	2.3 (1.6, 3.0)	2 (1, 3)	10.0 (5.8, 18.8)
1990	2033 (1510, 2634)	2.6 (1.9, 3.4)	2 (1, 4)	10.6 (5.5, 20.4)
2000	1704 (1297, 2198)	2.2 (1.6, 3.0)	2 (1, 4)	15.0 (7.4, 25.8)
Black women				
1980	28 (9, 127)	0.0 (0.0, 0.1)	0 (0, 0)	10.0 (5.8, 18.8)
1990	55 (20, 184)	0.0 (0.0, 0.2)	0 (0, 0)	10.6 (5.5, 20.4)
2000	54 (21, 173)	0.1 (0.0, 0.2)	0 (0, 0)	15.0 (7.4, 25.8)

counts within census tracts from the 1980, 1990 and 2000 US censuses for different age and race/ethnicity groups; and county-level population totals within age and race/ethnicity groups for the intercensal years (1981-1989 and 1991-1999).

Age at diagnosis is categorized into 8 groups: 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, and $84+$. We use internal standardization to calculate the expected number of breast cancer cases by CT for each time period. First, we ignore race/ethnicity and conduct our analysis for all women in LA county, standardizing the expected cases by age. Next, we restrict our analysis to the population of black women, again standardizing by age.

We quantify the socioeconomic gradient by calculating differences in breast cancer log-incidence rate ratios corresponding to an area-based socioeconomic measure (ABSM) between 1980 and 2000. We use the percent of the population below the poverty-level as the ABSM in our analyses.

Descriptive statistics for the LA cancer data are shown in Table 5.1. Figure 5.1 shows the county-level growth trends for the total and black populations. The number of census tracts in LA county was 1,633 in 1980; 1,642 in 1990 and 2,056 in 2000. The total population count of women over 15 years old was 2,993,192 in 1980; 3,492,249 in 1990; and 3,625,360 in 2000. The total number of black women over 15 years old was 369,543 in 1980; 402,207 in 1990; and 381,302 in 2000.



(a) Total women

(b) Black women

Figure 5.1: Population trends by age group in LA county for the female total and black populations. Age groups are shown above their respective trend lines.

5.3 Predicting intercensal population counts

The US Census provides intercensal population estimates within age and race/ethnicity groups at the county-level. However, in disease mapping studies, study investigators often use smaller geographic units, such as census tracts, as the units of analysis, to provide more geographic resolution in spatially heterogeneous populations. Census tracts (and block groups) are designed to be socioeconomically homogeneous; consequently, Krieger et al. (2002) report that socioeconomic health effects studies conducted at the census tract or block group level outperform zip-code aggregation. To calculate age-standardized expected breast cancer cases within census tracts, we need to estimate the population counts within age and race/ethnicity groups during the intercensal years (1981-1989, 1991-1999).

5.3.1 Boundary Normalization

Census tract boundaries shift over time at the census years, 1980, 1990, and 2000. In order to project intercensal population counts between the census years, we first need to normalize the census tract boundaries to a standard set of boundaries. The US census provides some normalized population counts, but does not break down these counts into age and race/ethnicity strata. Commercial GIS software provides normalized counts within these strata (Chen et al., 2008).

We normalize the 1980 and 2000 census tract boundaries to the 1990 boundaries. The primary challenge of normalizing the boundaries is that predictions must preserve Tobler's pyncophylactic property (Tobler, 1979). For instance, if two census tracts merge together, the count in the new tract should equal the sum of the counts in the two merged census tracts; if a census tract boundary did not change over time, then count should remain the same. Additionally, we need to ensure that the normalized predictions are positive.

The simplest normalization procedure that preserves these properties is proportional allocation of population counts (Gotway and Young, 2002). Gotway and Young (2007) propose an area-to-area kriging method to normalize boundaries that preserves the pyncophylactic property, but does not guarantee that projected counts are positive. We provide a review of proportional allocation and area-to-area kriging below.

We use both proportional allocation and area-to-area kriging to obtain two sets of normalized population counts. In our application, proportional allocation outperforms the kriging model in terms of root mean-square error (RMSE) and mean absolute prediction error (MAPE) in our validation set (see Section 5.4). In Section 5.3.2, we combine results from the kriging and proportional allocation models to produce more accurate intercensal interpolations using model stacking.

Area-to-area kriging and proportional allocation

In a given county, denote N_{ita} the population count for stratum a in census tract i at time t , denoted A_{it}. Denote the county-level population count in stratum a at time t by N_{ta}. The aggregate, county-level population counts $\{N_{ta}\}$ are known for all time points ($t = \{1, 2, ..., T\}$). The census tract population counts $\{N_{ita}\}$ are known only at census years ($t = 1, T$). In this example, we assume that we have counts from 2 census years, at time 1 and time T, but the method easily extends to more than two censuses.

Without loss of generality, we regularize population counts to the boundaries at time $t = 1$. First, we plot the empirical variogram using the centroids of the areas and determine a reasonable spatial model for interpolation. After finding an appropriate spatial covariance function for modeling the data, we predict census counts for the time $t = 1$ boundaries using the kriging equations Gotway and Young (2007):

$$N^*_{jTa} = \sum_i w_{ija} N_{ita}$$

where N^*_{jTa} are the estimated counts at time T in census tract j, after changing to the time 1 boundaries. The kriging weights $\boldsymbol{w_{ja}} = (w_{1ja}, ..., w_{K_i ja})$ are:

$$\boldsymbol{w_{ja}} = \boldsymbol{\Sigma_{T^*T,a}} \boldsymbol{\Sigma^{-1}_{TT,a}}$$

where $\boldsymbol{\Sigma_{T^*T,a}}[1, i] = Cov(N^*_{jTa}, N_{iTa})$ is a $1 \times i$ matrix describing the covariance between the counts at the time 1 and time T boundaries; and $\boldsymbol{\Sigma_{TT,a}}[i, j] = Cov(N_{jTa}, N_{iTa})$ is a $i \times i$ variance-covariance matrix for the counts at time T. Note that $\sum_i |A_{it}| w_{ija} = |A_{jT}|$.

While our variogram fitting model is rather ad-hoc, we do not believe that the form of the variogram will be as important in area-to-area kriging, as both the area of the census tracts and the spatial covariance function play a role in determining the new interpolated counts. We could use the more sophisticated variogram fitting methods presented in Gotway and Young (2007). In our application, the number of areas is large and the population counts are variable, which can produce unstable kriging results (*e.g.* negative population counts in areas). Following Gotway and Young (2007), we use a local kriging approach to

predict the census tract counts, using only the nearest 5-15 neighboring areas to predict the new count.

If the spatial structure of the population counts is ignored and proportional allocation is used to normalize boundaries, then $w_{ija} = |A_{ijt}|/|A_{it}|$, where $|A_{ijt}|$ is the area of the portion of census tract i at time t that is contained in area j at time T. Again, $\sum_i |A_{it}| w_{ija} = |A_{jT}|$.

5.3.2 Interpolating intercensal counts

After obtaining normalized census tract counts in 1980 and 2000, we interpolate intercensal tract counts using the 1990 census boundaries as our reference set of boundaries. We predict census tract population counts within age and race/ethnicity strata at all intercensal years (1981-1989, 1991-1999), given (1) census tract population counts within age and race/ethnicity strata at the census years (1980, 1990, 2000) and (2) county-level population counts within age and race/ethnicity strata at all years (1980-2000). In this section, we propose multiple models for intercensal population count interpolation; these models are also listed in Section 5.3.3.

Linear Interpolation

First, we interpolate using simple linear interpolation, assuming that the population within a census tract changes linearly over time. Linear interpolation does not require the county-level population totals, and therefore ignores any county-level population growth trends.

Apportionment Probabilities

Best and Wakefield (1999) interpolate intercensal counts via apportionment probabilities. In our application, the apportionment probability p_{ita} is the fraction of the population

within stratum a in LA county living in census tract i at time t, $p_{ita} = N_{ita}/N_{ta}$. Apportionment probabilities are known at the census years and are extrapolated to the intercensal years. To impute intercensal counts, apportionment probabilities are modelled using a logistic-linear model:

$$\text{logit}(p_{ita}) = \beta_{0ia} + \beta_{i1a}t$$

Then, $\hat{N}_{ita} = N_{ta}p_{ita}$.

This model does not impose any standard population growth or decay model on the population counts. Additionally, $\sum_i p_{it} \neq 1$, so we adjust the apportionment probabilities to sum to 1.

In addition to the linear-logistic model, we also fit a model assuming that apportionment probabilities increase linearly over time within census tracts.

Additive model for population change over time

Next, we construct an additive model for the intercensal population counts. We estimate changes in the population within a stratum over time, modelling $N_{ita} = N_{i(t-1)a} + r_{ita}$, where r_{ita} is the unknown population growth parameter of interest. Because data reflecting births, deaths, or migration at the census tract level within strata are not available, we cannot incorporate any direct population growth or decline data into our model. Consequently, we build a model based on the concept that populations change gradually over time. Events such as Hurricane Katrina are of course exceptions to this rule, but, in most circumstances, this model should be reasonable.

We assume population changes are smooth over time within each census tract and meet the county-level total population constraints at each year. Specifically, the parameters $\{r_{ita}\}$ encompass population growth or decline due to births, deaths, and migration. We solve for $\{r_{ita}\}$ by minimizing an objective function that imposes smoothness con-

straints on the population growth parameters r_{ita}, *e.g.*

$$\sum_{i=1}^{n}\sum_{t=1}^{T-1}|N_{iTa}-N_{i1a}|^{-q}r_{ita}^2.$$

In this model, q is a user-specified parameter that adjusts smoothness in the growth parameters $\{r_{ita}\}$ by the total amount of growth in a census tract over time. Specifically, choosing $q = 0$, areas experiencing a lot of growth over time will have very smooth growth trends over time, whereas those with less growth may have more erratic growth patterns. Choosing $q = 1$ would provide more balance between areas with different net population growth over time.

We have $n(t-1)$ unknown parameters, $\{r_{ita}\}$. Without imposing any constraints, the solution to the minimization problem is $\{r_{ita}\} = \mathbf{0}$. However, we aim to minimize this quadratic objective function, subject to the three linear constraints: (1) $\sum_t r_{ita} = N_{iTa} - N_{i1a}$, (2) $\sum_i r_{ita} = N_{ta} - N_{(t-1)a}$, and (3) $N_{ita} > 0$.

We solve the constrained optimization problem using the `quadprog` function in `Matlab`. Using this additive growth model, the parameters $\{r_{ita}\}$ are directly interpretable, and we can compare many different models by changing the objective function.

We avoid framing the intercensal population count imputation problem within a probabilistic framework, *e.g.* modelling the population counts using a Poisson distribution. Determining a parametric distribution for the population counts is difficult, given the skewed distribution of census tract counts within strata and the linear constraints. Further, we are more concerned with interpolation, and using a probabilistic approach can result in oversmoothing. Lastly, if we knew the correct model for the intercensal population counts, then we would know the exact intercensal population count (because census tract counts are known at the census years). For instance, if we knew the birth, death, and migration rates for each substratum in each census tract, then we would know the intercensal population counts. Given that we do not know the correct model, the error that we are concerned with is model misspecification, rather than sampling error. This distinguishes our approach from the probabilistic framework outlined in Best and

Wakefield (1999), described in Section 5.7.1.

Model stacking

Lastly, we can consider a linear combination of the intercensal models. In our disease mapping model, we aim to estimate the expected number of breast cancer cases in a census tract as accurately as possible in the intercensal years. To determine the optimal weights for the linear combination, we use the 1980 and 2000 census data to predict the 1990 census tract expected case counts for each model and compare the predicted expected counts to the actual 1990 census data.

Denote E_{it} as the expected breast cancer case count in census tract i at time t, and $\hat{E}_{itm}$ as the estimated expected breast cancer case count using intercensal model m. We fit K different intercensal population models, $\{M_1, ..., M_K\}$. Let ω_m denote the weight assigned to model m, $\sum_m \omega_m = 1$. (We omit the time index on ω_m because we are only predicting the expected counts at one year, 1990.) We construct a linear combination of the models, $\hat{E}^S_{it} = \sum_m \omega_m \hat{E}_{itm}$ and estimate ω_m using model stacking (Wolpert, 1992). Traditional stacking minimizes

$$\sum_{i=1}^{n} \left(E_{it} - \sum_{m=1}^{K} \omega_m \hat{E}_{itm} \right)^2$$

with respect to $\{\omega_m\}$ to obtain model weights. The distribution of census tract counts within strata is highly skewed, and the L-2 norm will emphasize the prediction of intercensal counts in tracts with very large populations. Therefore, to avoid outliers driving the choice of the weights, we also estimate the set of weights $\{\omega^1_m\}$ that minimizes the objective function with respect to the L-1 norm,

$$\sum_{i=1}^{n} \left| (E_{it} - \sum_{m=1}^{K} \omega^1_m \hat{E}_{itm} \right|.$$

Model stacking weights for the L-1 and L-2 norm are shown in Table 5.2. Traditional Bayesian model averaging weights (Hoeting et al., 1999) are also shown in Table 5.2, to contrast the stacking weights to model averaging weights. Stacking selects the best linear

Table 5.2: Stacking weights for L-1 and L-2 norm; and Bayesian model averaging weights, in the total and black populations.

		Total Population			Black Population		
		Stacking			Stacking		
	Model	L-1	L-2	BMA	L-1	L-2	BMA
Prop. alloc.	A1	0.063	0.121	0.252	0.099	0.000	0.000
	A2	0.000	0.000	0.000	0.002	0.000	0.000
	A3	0.000	0.000	0.000	0.126	0.501	0.000
	LI	0.855	0.789	0.748	0.180	0.000	0.000
	P1	0.000	0.000	0.000	0.478	0.213	0.000
	P2	0.000	0.000	0.000	0.000	0.000	1.000
Krige	A1	0.000	0.000	0.000	0.000	0.000	0.000
	A2	0.000	0.000	0.000	0.000	0.000	0.000
	A3	0.000	0.000	0.000	0.000	0.000	0.000
	LI	0.082	0.037	0.000	0.000	0.000	0.000
	P1	0.000	0.053	0.000	0.114	0.286	0.000
	P2	0.000	0.000	0.000	0.000	0.000	0.000

combination of the fitted models to improve prediction, whereas Bayesian model averaging assumes the data generating model is in the set of candidate models and weights the models accordingly to incorporate model uncertainty.

Intercensal models

We fit various different intercensal models in our analysis, including linear interpolation; two apportionment probability models (linear and logistic-linear); three additive population growth models ($q = 0, 0.5, 1$); and the two stacked models (L-1 and L-2 norm). In Section 5.3.3, we list the proposed intercensal models (and the abbreviations that we use for these models). We fit each model using proportional allocation and area-to-area kriging for boundary normalization. In Figure 5.2, we plot two sample trajectories of expected counts from the different models. We observe differences in the expected counts at the census years in the linear interpolation model due to the fact that we use internal standardization, and we did not constrain the sum of the population counts at the intercensal years to match the county-level population.

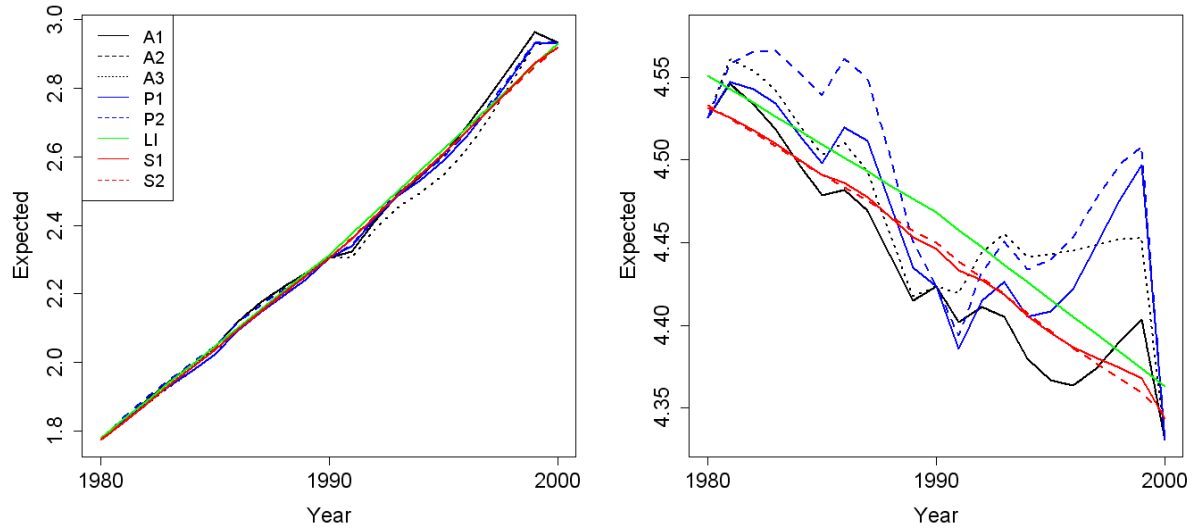


Figure 5.2: Intercensal expected count estimates for the different models. Panel 1 shows a census tract with a substantial amount of population change over time, and panel 2 shows a census tract with very little change over the 20 year time period. Expected counts were calculated using proportional allocation for boundary normalization.

5.3.3 Final list of intercensal interpolation models.

In this Section, we list the intercensal models fit in the validation procedure and analyses in the remainder of the paper.

1. Additive models, for $q = 0, 0.5, 1$ (A1, A2, A3), minimize

$$\sum_{i=1}^{n}\sum_{t=1}^{T} |N_{iTa} - N_{i1a}|^{-q} r_{ita}^2$$

   subject to linear constraints in Section 5.3.2.

2. Apportionment probabilities - $N_{ita} = N_{ta}p_{ita}$, where

   - Logistic-linear (P1): $logit(p_{ita}) = \beta_0 + \beta_1 t$
   - Linear (P2): $p_{ita} = \beta_0 + \beta_1 t$

3. Linear interpolation (LI): $N_{ita} = N_{i(t-1)a} + \alpha_{i(t-1)a}$, $\alpha_{ita} = \frac{N_{Ta} - N_{1a}}{T-1}$.

We fit each of the 6 different intercensal interpolation models (A1, A2, A3, P1, P2, L1) to the normalized census counts calculated using area-to-area kriging and using proportional allocation. Lastly, we combine the 12 models above using model stacking with the L-1 and L-2 norms (models S1 and S2, respectively). Therefore, we fit 14 unique intercensal population count models.

5.4 Simulation Study

To gauge the impact of denominator uncertainty in our disease mapping model, we use the 1980 and 2000 censuses and the county-level population totals from 1980-2000 to predict the expect breast cancer counts in 1990. We estimate the expected breast cancer case counts between 1981 and 1999, using the 1980 and 2000 census tract data and the county-level total populations at each time point. We then compare the predicted expected counts for each model $\hat{E}_{i(1990)m}$ to the true expected counts in 1990, $E_{i(1990)}$. We use internal standardization to calculate the expected counts, with age-specific probabilities calculated using only breast cancer cases and true census tract counts from 1990.

To determine which model for the expected case counts has the lowest prediction error, we compare the root mean squared error, $RMSE = (1/n)(\sum_{i=1}^{n}(E_{it} - \hat{E}_{itm})^2)^{-1/2}$; and mean absolute prediction error, $MAPE = (1/n)\sum_{i=1}^{n}|E_{it} - \hat{E}_{itm}|$. Comparing predictions in the expected counts is preferable to comparing absolute population counts, because the final disease mapping model relies on the population counts only through the expected case counts. Therefore, we should favor models that produce estimates of the expected counts that are closest to the expected counts calculated from the census data.

Next, we perform a simulation study to assess whether errors in denominators can induce bias in coefficient estimates. We generate simulated datasets by using the expected cancer case counts and percent of the population below poverty in the census tract. Expected case counts are again calculated using internal standardization, but now use data

from all three census years. We generate outcomes $Y_{it} \sim Pois(\mu_{it})$, where

$$log(\mu_{it}) = log(E_{it}) + \beta_0 + \beta_1 x_{it} + \beta_2 t + \beta_3 x_{it} t$$

where x_{it} is the poverty indicator in census tract i at time t, $t = \{0, 10, 20\}$, and $\beta_0 = 0.1, \beta_1 = -1, \beta_2 = 0$, and $\beta_3 = 0.025$.

After generating 1,000 simulated datasets using the 1990 census data, we fit the above model to each simulated dataset, *changing only the expected case counts to the predicted counts*. Then, we compare estimates of β_3 (analogous to the socioeconomic gradient) across the models. We also fit a model using only the 1990 data to assess bias in the estimate of the association between socioeconomic status and breast cancer incidence; specifically, we model the linear predictor as $log(\mu_{it}) = log(E_{it}) + \alpha_0 + \alpha_1 x_{it}$, where $t = 10$, and assess bias in α_1 (note that $\alpha_1 = -0.75$ under the data generating model).

We fit the two generalized linear models using the estimated expected counts $\{\hat{E}_{itm}\}$ for each set of interpolated census counts m. By changing only the denominators, we can assess the impact that denominator uncertainty has on estimating the association between socioeconomic status and breast cancer incidence and changes in the socioeconomic gradient over time in our disease mapping study. We compare 14 different models, listed in Section 5.3.3; using this comprehensive list, we can compare the performance of proportional allocation versus kriging; and the relative performance of the intercensal interpolation models in Section 5.3.2.

Results from the simulation study are presented in Table 5.3. Prediction errors are generally similar between all of the models, and the stacked models have the lowest prediction error, as expected. The prediction error is consistently lower in the models using proportional allocation for boundary normalization (versus area-to-area kriging).

Examining the total population, bias in the estimates of β_3 and α_1 is negligible in the majority of the models. Substantial bias is observed only in the linear-logistic apportionment probability model.

When we restrict to the black population only, bias in the estimates of β_3 and α_1 is

Table 5.3: Prediction error and bias (percent bias) in estimating the effect of SES on breast cancer incidence and changes in the relationship over time. Census denotes the model using the 1990 census data, which is the data generating model in the simulation study. The true SES effect is $\alpha_1 = -0.75$ and true interaction coefficient is $\beta_1 = .025$. The first column denotes whether the model uses proportional allocation (PA) or kriging (K) for normalization of boundaries; the last two rows are the stacked models (S).

		Total Population						Black Population					
		Prediction err.		SES effect		Interaction		Prediction err.		SES effect		Interaction	
	Model	RMSE	MAPE	$E(\hat{\alpha}_1)$	%	$E(\hat{\beta}_3)$	%	RMSE	MAPE	$E(\hat{\alpha}_1)$	%	$E(\hat{\beta}_3)$	%
	Truth	0	0	-0.74	0.7	0.026	3.2	0	0	-0.76	-1.9	0.025	-1.4
P	A1	.313	.192	-0.76	-1.1	.024	-2.3	.089	.045	-0.26	65.9	.075	202.0
	A2	.317	.194	-0.76	-1.5	.024	-3.6	.086	.042	-0.34	54.9	.067	168.9
	A3	.332	.202	-0.77	-3.3	.023	-8.9	.083	.040	-0.42	44.3	.059	137.3
	LI	.313	.191	-0.75	-0.1	.025	0.5	.090	.041	-0.51	32.1	.050	100.5
	P1	.409	.229	-0.91	-21.5	.009	-63.4	.082	.035	-1.15	-53.4	-.014	-156.0
	P2	.335	.201	-0.80	-6.6	.020	-18.9	.078	.034	-0.77	-2.7	.024	-3.9
K	A1	.411	.226	-0.71	5.3	.029	16.9	.094	.046	-0.28	62.1	.073	190.5
	A2	.412	.228	-0.73	3.1	.028	10.4	.091	.043	-0.36	51.5	.065	158.8
	A3	.422	.234	-0.75	-0.1	.025	0.5	.089	.041	-0.44	41.4	.057	128.5
	LI	.411	.223	-0.72	4.6	.029	14.9	.095	.042	-0.52	30.1	.049	94.5
	P1	.451	.251	-0.90	-20.0	.010	-59.2	.089	.038	-1.19	-58.1	-.018	-170.1
	P2	.426	.234	-0.76	-1.5	.024	-3.4	.084	.037	-0.77	-2.8	.024	-4.2
S	S1	.312	.191	-0.75	0.2	.025	1.6	.069	.029	-0.85	-13.4	.016	-35.8
	S2	.311	.191	-0.76	-1.1	.024	-2.4	.067	.029	-0.78	-4.5	.023	-9.2

Table 5.4: Relationship between the bias in interpolation and SES; and the relationship between the true expected count and the bias in the expected counts. Coefficients represent percent increase in poverty in a census tract for a 0.1 unit increase in bias; and the increase in bias for a 1 unit increase in expected count. The first column denotes whether the model uses proportional allocation (P) or kriging (K) for normalization of boundaries; the last two rows are the stacked models (S).

		Total Population						Black Population					
		$SES \sim (\hat{E} - E)$			$(\hat{E} - E) \sim E$			$SES \sim (\hat{E} - E)$			$(\hat{E} - E) \sim E$		
	Model	Coef.	SE	Z	Coef.	SE	Z	Coef.	SE	Z	Coef.	SE	Z
P	A1	-0.05	0.09	-0.5	0.04	0.01	6.7	2.63	0.32	8.2	0.10	0.00	35.2
	A2	-0.05	0.09	-0.5	0.04	0.01	5.9	2.32	0.34	6.9	0.08	0.00	28.8
	A3	-0.08	0.09	-0.9	0.04	0.01	5.6	1.99	0.35	5.7	0.06	0.00	21.3
	LI	-0.04	0.09	-0.4	0.04	0.01	6.3	2.63	0.32	8.2	0.10	0.00	35.1
	P1	-0.26	0.07	-3.7	0.05	0.01	6.7	-2.45	0.35	-7.0	-0.03	0.00	-10.3
	P2	-0.13	0.09	-1.6	0.04	0.01	5.9	-0.26	0.37	-0.7	0.04	0.00	11.4
K	A1	0.04	0.07	0.6	0.04	0.01	5.0	2.22	0.31	7.2	0.10	0.00	32.5
	A2	0.02	0.07	0.3	0.04	0.01	5.0	1.93	0.32	6.1	0.08	0.00	26.9
	A3	-0.02	0.07	-0.3	0.05	0.01	5.6	1.63	0.33	5.0	0.07	0.00	20.5
	LI	0.05	0.07	0.7	0.04	0.01	4.6	2.22	0.31	7.2	0.10	0.00	32.3
	P1	-0.21	0.06	-3.2	0.05	0.01	5.8	-2.23	0.32	-6.9	-0.03	0.00	-9.3
	P2	-0.04	0.07	-0.5	0.05	0.01	5.7	-0.22	0.35	-0.7	0.04	0.00	11.9
S	S1	-0.03	0.09	-0.28	0.04	0.01	6.29	-0.52	0.42	-1.22	0.02	0.00	5.48
	S2	-0.05	0.09	-0.57	0.04	0.01	6.45	-0.39	0.43	-0.91	0.02	0.00	5.68

low only in the linear apportionment probability model and in the stacking model using the L-2 norm. In the remaining models, bias in the coefficient and interaction estimates is high.

We examine the source of the bias in these estimates in Table 5.4. First, for each model m, we regress $E_{it} - \hat{E}_{itm}$ on E_{it}, to assess the relationship between the true expected cancer counts and prediction error in the estimates. In census tracts with higher expected counts, prediction error tends to be greater.

Next, we regress x_{it} (the percent of the population below poverty in a census tract) on the prediction error, $E_{it} - \hat{E}_{itm}$. In the total population, poverty is associated with prediction error only in the two models with substantial bias. When we restrict to the black population only, poverty is correlated with the prediction error.

Specifically, an increase in the difference between the true and predicted expected breast cancer counts $E_{it} - \hat{E}_{itm}$ is strongly associated with an increase in the poverty indicator. In most of the interpolation models, predicted expected counts were overestimated in census tracts with more poverty, resulting in substantial underestimation of the relationship between SES and breast cancer incidence in 1990 and consequently an overestimate in the socioeconomic gradient. In summary, we found that correlation between the covariate of interest and expected counts can result in systematic bias in fixed effects estimates in disease mapping models, particularly when stratifying by race/ethnicity.

In the above simulation study, we only used data from the three census years, so that we could compare the true census data with the predicted counts to gauge the impact of bias. However, when fitting a disease mapping model to twenty time points, rather than just three, we might observe less bias. Therefore, we conducted additional simulations to assess the impact of denominator uncertainty across multiple time points. We selected three models to use as our "true" intercensal population counts (i.e. as our data generating model): (1) the linear apportionment probability model, (2) the additive model with $q = 0$, and (3) the stacked model using the L-2 norm. We used only the models with

proportional allocation boundary normalization. After generating the data using one of the three models, we fit the same model, changing only the expected counts (and using those from remaining intercensal count models), and assessed bias in the estimates of β_3. Results of the simulation study are shown in Table 5.5.

Using data from the twenty different time points, we find that estimates of the socioeconomic gradient are similar across all models, especially in the total population. We still observe some bias in $\hat{\beta}_3$ for many of the models, but bias is greatly attenuated compared to the simulations that use data from only the three census years. The estimates of the association between the poverty indicator and incidence are biased, but the bias is similar across time points, resulting in only moderate bias in the estimate of the socioeconomic gradient (results not shown).

The stacking model performed reasonably well in terms of coefficient bias, and we use the L-2 norm stacked model in our final analyses.

5.5 General Spatial Misalignment Framework

Now that we have population denominator data for each time point, we can estimate the association between SES and breast cancer incidence between 1980 and 2000. In order construct this spatio-temporal disease mapping model, we need a plausible model for residual spatial variability in the data. Area-to-area spatial misalignment arises in our data, because of the boundary differences in the 1980, 1990, and 2000 census tracts. We could use the normalized census tract counts in 1980 and 2000 (normalized to the 1990 boundaries), but this solution would introduce unnecessary uncertainty into our model. Handling area-level temporal misalignment is simple using Hund et al. (2012), but flexibility in the structure of the residual spatio-temporal variability is limited.

Table 5.5: Heterogeneity in estimates of the socioeconomic gradient, $\hat{\beta}_3$ due to differences in intercensal count interpolation models, when all 21 years of data are used. Estimates of β_3 and percent bias in the estimate are show. Three different data generating models are used, and are denoted with a * in the table.

		Total Population						Black Population					
Model		$E(\hat{\beta}_3)$	%	$E(\hat{\beta}_3)$	%	$E(\hat{\beta}_3)$	%	$E(\hat{\beta}_3)$	%	$E(\hat{\beta}_3)$	%	$E(\hat{\beta}_3)$	%
P	A1	.025	-1.1	$.025^*$	0.9	.025	-0.8	.033	33.0	$.026^*$	2.2	.034	34.4
	A2	.025	-0.3	.025	1.7	.025	0.0	.029	17.8	.022	-13.5	.030	19.2
	A3	.025	0.1	.026	2.2	.025	0.5	.025	1.6	.017	-30.3	.026	3.0
	LI	.026	3.9	.026	6.0	.026	4.2	.019	-22.0	.011	-54.6	.020	-20.6
	P1	.027	7.5	.027	9.6	.027	7.9	.028	11.7	.020	-20.9	.028	13.1
	P2	$.025^*$	-0.4	.025	1.6	.025	-0.1	$.025^*$	-0.7	.017	-33.1	.025	0.8
K	A1	.024	-5.2	.024	-3.3	.024	-4.4	.032	28.5	.025	-1.9	.032	29.5
	A2	.024	-2.5	.025	-0.6	.025	-1.7	.028	12.3	.020	-18.6	.028	13.4
	A3	.025	-1.0	.025	0.9	.025	-0.1	.024	-4.1	.016	-35.5	.024	-3.0
	LI	.025	-0.4	.025	1.5	.025	0.5	.016	-34.8	.008	-67.1	.017	-33.7
	P1	.026	2.1	.026	4.0	.026	3.0	.025	-1.9	.016	-34.4	.025	-0.8
	P2	.024	-4.3	.024	-2.4	.024	-3.4	.021	-14.2	.013	-46.3	.022	-13.1
S	S1	.025	-0.1	.025	2.0	.025	0.2	.025	-0.4	.017	-32.9	.025	1.0
	S2	.025	-0.0	.026	2.0	$.025^*$	0.3	.024	-4.6	.016	-37.3	$.024^*$	-3.1

5.5.1 Review of Area-level Geoadditive Models

Banerjee et al. (2008) and Kammann and Wand (2003) discuss reduced rank geoadditive modelling for point-level data. Hund et al. (2012) propose a geoadditive model for misaligned area-level data, constructed within the generalized linear mixed model framework. To bypass issues with temporal data misalignment, they model the underlying continuous spatial surface and aggregate to the area-level by using a quadrature approximation. In the section below, we extend the Hund et al. (2012) to allow for more general spatio-temporal variability.

For a collection of disjoint areas $\boldsymbol{s}$ in the study region $\boldsymbol{A}$, we model the outcome $\boldsymbol{Y}(\boldsymbol{s})$ and area-level covariates $\boldsymbol{X}(\boldsymbol{s})$ using an exponential family, $\boldsymbol{Y}(\boldsymbol{s}) \sim \mathrm{f}(\boldsymbol{\mu}, \boldsymbol{\theta})$, where the linear predictor $\boldsymbol{\eta}(\boldsymbol{s})$, as a function of covariates and spatial random effects. We select a fine grid of evenly-spaced quadrature points $\boldsymbol{\omega}$ over the study region, and model the underlying latent continuous spatial process. We aggregate over the grid of quadrature points to obtain the area-level model:

$$\boldsymbol{\eta}(\boldsymbol{a}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u},$$

where $\boldsymbol{W}(\boldsymbol{s})$ is a quadrature weight matrix that aggregates the latent continuous surface over each area to obtain area-level random effects; the matrix $\boldsymbol{Z}(\boldsymbol{a})$ are basis functions projecting from the quadrature points to a set of fixed knot locations using a spatial correlation function $C(\boldsymbol{\theta})$; $\boldsymbol{u}$ are basis coefficients estimated via model fitting.

To define $\boldsymbol{Z}(\boldsymbol{a})$, we choose a set of knots $\{\kappa_g\}_{g=1,\ldots,G}$. The $G \times G$ matrix $\boldsymbol{\Omega} = [C(|\kappa_i - \kappa_j|)]_{i,j=1}^{G}$ is a variance-covariance matrix modelling spatial variability between the knots. The $M \times G$ matrix $\tilde{\boldsymbol{Z}}(\boldsymbol{s}) = [C(|\boldsymbol{s} - \kappa_j|)]_{j=1}$ projects from the quadrature point locations to the knot grid. Then, we construct $\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{Z}\tilde{(\boldsymbol{s})}\boldsymbol{\Omega}^{-1/2}$, and finish specification of the model by assuming $\boldsymbol{u} \sim MVN(0, \sigma^2\boldsymbol{I})$. Note that the predictive process model of Banerjee et al. (2008) defines $\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{Z}\tilde{(\boldsymbol{s})}\boldsymbol{\Omega}^{-1}$ and $\boldsymbol{u} \sim MVN\{0, \sigma^2\boldsymbol{C}(\boldsymbol{\theta})\}$. These models produce very similar results, but we use the Kammann and Wand (2003) model to facilitate computationally efficient frequentist model fitting in standard software. Choosing

$\boldsymbol{W}(\boldsymbol{s}) = \boldsymbol{I}$, we arrive at the geoadditive point-level model of Kammann and Wand (2003).

Note that $\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{u}$ is not directly interpretable in terms of the area-level residual spatial relative risk unless $\mathrm{f}(\cdot)$ is the identity link function (*e.g.* when the outcome is normally distributed).

5.5.2 Multivariate spatial regression with misalignment

To construct a multivariate spatial regression model that bypasses model fitting difficulties associated with spatial misalignment, we exploit two important features of the geoadditive framework described above: (1) the area-level model is a generalization of the point-level model; and (2) spatial variability is induced though a fixed set of knots.

Suppose we now have J outcome types $\boldsymbol{Y_j}(\boldsymbol{s_j})$ observed over $\boldsymbol{s_j}$ locations or areas. Consider the following multivariate model for linear predictor $\boldsymbol{\eta_j} = \boldsymbol{g}(\boldsymbol{\mu_j})$:

$$\boldsymbol{\eta_j}(\boldsymbol{s_j}) = \boldsymbol{X_j}\boldsymbol{\beta_j} + \boldsymbol{Z_j}(\boldsymbol{s_j})\boldsymbol{u_j}$$

where i indexes individual and j indexes type of outcome. The basis coefficients $\boldsymbol{Z_j}(\boldsymbol{s_j})$ are constructed based on a fixed set of knots, identical to the coefficients in the univariate setting. The form of the coefficients $\boldsymbol{Z_j}(\boldsymbol{s_j})$ depends on whether the outcome is area- or point-referenced. However, it is clear that model-fitting does not depend on point- or area-level alignment of data locations, and mixing area- and point-level outcomes is consequently trivial.

In the multivariate model, we induce correlation between the random effects at the knot-locations to induce correlation between the spatial surfaces for each outcome type. Specifically, $\boldsymbol{u_j} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a $J \times J$ unstructured variance-covariance matrix. The random effects at different knot locations, e.g. $\boldsymbol{u}_l$ and $\boldsymbol{u}_k$, are independent. Note that the parametric form for $\boldsymbol{\Sigma}$ is flexible, and, as an example, and we could replace the unstructured matrix with an AR(1) covariance matrix if we were working with spatio-temporal data. Hence, the model affords a fair amount of flexibility regarding choice

of the covariance structure between outcomes. Additional random effects (for instance, within-subject correlation), can also be added to the model, when relevant.

5.5.3 Spatial Confounding

Because poverty in LA county varies spatially, spatial confounding induced by collinearity between the spatial random effects and poverty indicator could distort the relationship between poverty and breast cancer incidence in our disease mapping study (Hodges and Reich, 2010). To avoid bias in estimation of our fixed effects due to this collinearity, we project $\boldsymbol{Z_t}(\boldsymbol{s})$ at each time t, following Reich et al. (2006). The projected basis functions are defined as $\boldsymbol{Z_t^P}(\boldsymbol{s}) = \boldsymbol{P_t}\boldsymbol{Z_t}(\boldsymbol{s})$, where $\boldsymbol{P_t} = \boldsymbol{I} - \boldsymbol{X_t}(\boldsymbol{X_t'}\boldsymbol{X_t})\boldsymbol{X_t'}$.

5.5.4 Model fitting

We fit our disease mapping model using the PQL approximation to maximum likelihood (Breslow and Clayton, 1993). Alternatively, we could use the computationally efficient Bayesian predictive process model, noting that the distribution of the random effects and definition of the random effects design matrix would be slightly different (Banerjee et al., 2008). As in Banerjee et al. (2008), we can use the Sherman-Woodbury matrix inversion formulas (Harville, 2008) to improve the computational efficiency of the PQL estimation algorithm, inverting a $JG \times JG$ matrix, rather than an $M \times M$ matrix, where M is the total sample size.

Estimation of the variance components, $\boldsymbol{\Sigma}$, is the most difficult aspect of the model fitting procedure. In our disease mapping model, we assume $\boldsymbol{\Sigma}$ has a heterogeneous AR1 structure, namely $Cov(u_{it}, u_{jt}) = 0, Var(u_{it}) = \sigma_t^2$, and $Cov(u_{it}, u_{it'}) = \sigma_t\sigma_{t'}\rho^{|t-t'|}$. Following Hund et al. (2012), we model spatial variability using an exponential variance structure, $C(s_i, s_j) = \sigma_t^2 exp(-r|s_i - s_j|)$, and fix the range parameter r. Hence, we have 22 variance components in our disease mapping model.

When the number of variance components is large, maximizing the profile log-likelihood becomes increasingly computationally intensive. We consider an alternative computationally efficient algorithm for estimating the variance components. First, we fit univariate disease mapping models at each time point, and estimate the marginal variance components, $\{\sigma_t^2\}$. Then, we maximize the profile likelihood to estimate ρ, fixing the marginal variance components.

5.6 Data Application

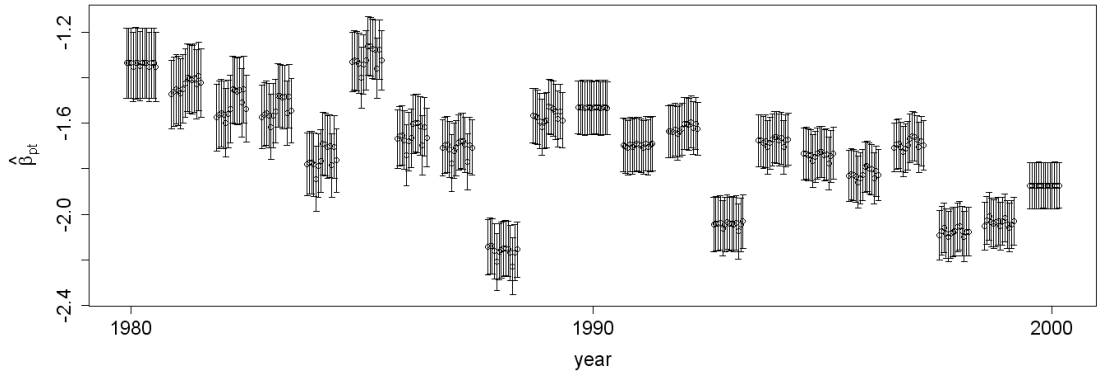
To examine the impact of denominator uncertainty in our analysis, we first fit a Poisson generalized linear model to the data, ignoring any residual spatial variability. Let Y_{it} denote the observed number of incident breast cancer cases in CT i at time t, and assume $Y_{it} \sim Poisson(\mu_{it})$. We fit the model:

$$\log(\mu_{it}) = \log(\hat{E}_{itm}) + \beta_t + \beta_{pt} x_{it}$$

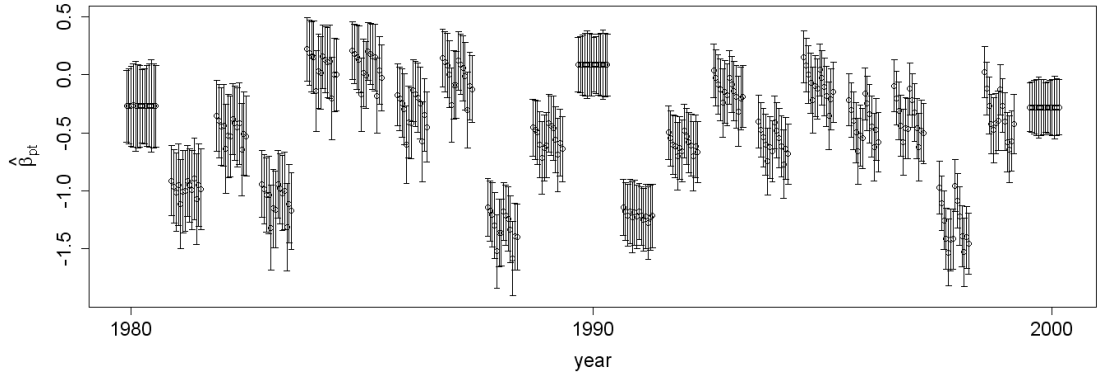
where x_{it} is the percent of the population below poverty in area i at time t. We introduce an overdispersion parameter ϕ into the model to account for additional non-spatial variability in the data greater than that predicted by the Poisson distribution.

We fit the model for each set of expected counts described in Sections 5.3 and 5.4. In Figure 5.3, we plot $\hat{\beta}_{pt}$ over time for each set of counts in both the total and black populations.

Paralleling our simulation study, estimates of the socioeconomic gradient in breast cancer incidence in the total population are similar across all intercensal population interpolation models, even though estimates of the relationship between SES and breast cancer incidence at each time point differ between models. Restricting to black women only, we observe more heterogeneity in estimates of the socioeconomic gradient, as well as in the estimates of the relationship between SES and breast cancer incidence at each time point, due to the sparsity of breast cancer cases and the highly skewed distribution of the population denominators.



(a) Total population



(b) Black population

Figure 5.3: Estimates of $\hat{\beta}_{pt}$ and corresponding 50% confidence intervals for the different intercensal count models.

In the total population, we observe an apparent decrease in $\hat{\beta}_{pt}$ over time. To quantify this decrease, we fit a more parametric model, assuming breast cancer incidence and the socioeconomic gradient change linearly over time,

$$\log(\mu_{it}) = \log(\hat{E}_{itm}) + \beta_0 + \beta^T t + \beta^P x_{it} + \beta^{PT} x_{it} t,$$

for each model m. We compare estimates of $\hat{\beta}^{PT}$ for each model in the total and black populations in Table 5.6. Ignoring residual spatial variability, we find that the socioeconomic gradient in breast cancer incidence in the total population appears to be increasing over time ($p < 0.001$). We observe no change in the socioeconomic gradient in breast cancer incidence over time in the black population.

Table 5.6: Estimates of $\hat{\beta}^{PT}$ for the intercensal models.

		Total Population			Black Population		
Model		$\hat{\beta}_{pt}$	$se(\hat{\beta}_{pt})$	p-value	$\hat{\beta}_{pt}$	$se(\hat{\beta}_{pt})$	p-value
P	A1	-0.025	0.007	0.0002	0.008	0.014	0.552
	A2	-0.024	0.007	0.0002	0.004	0.014	0.770
	A3	-0.024	0.007	0.0002	-0.000	0.015	0.997
	LI	-0.024	0.007	0.0004	-0.006	0.016	0.680
	P1	-0.023	0.007	0.0008	0.002	0.017	0.900
	P2	-0.025	0.007	0.0002	-0.001	0.016	0.942
K	A1	-0.028	0.007	< 0.0001	0.007	0.014	0.625
	A2	-0.027	0.007	< 0.0001	0.003	0.014	0.860
	A3	-0.027	0.007	< 0.0001	-0.002	0.015	0.911
	LI	-0.027	0.007	< 0.0001	-0.010	0.016	0.531
	P1	-0.025	0.007	< 0.0001	-0.001	0.018	0.973
	P2	-0.028	0.007	< 0.0001	-0.005	0.016	0.770
S	S1	-0.025	0.007	0.0002	-0.001	0.016	0.948
	S2	-0.025	0.007	0.0002	-0.002	0.016	0.895

To address the potential impact of residual spatial variability in our analysis, we fit the geoadditive disease mapping model, modeling the linear predictor,

$$log(\mu_{it}) = \beta_t + \beta_{pt} x_{it} + \boldsymbol{Z_{it}(s_{it})u_t}$$

where the basis functions and coefficients $\boldsymbol{Z_{it}(s_{it})}$ and $\boldsymbol{u_t}$ are defined as in Section 5.5. We also fit the projected model to bypass issues with spatial confounding, defining the basis functions as $\boldsymbol{Z^P_{it}(s_{it})}$ (as in Section 5.5.3).

To model spatial variability, we use the heterogeneous AR(1) correlation structure for the distribution of the basis coefficients $\boldsymbol{u_t}$ (Section 5.5.4), and construct the basis functions $\boldsymbol{Z_{it}(s_{it})}$ using the exponential correlation structure, with $Corr(s_{it} - \kappa_g) = exp(-|s_{it} = \kappa_g|\nu)$. We choose $\nu = 15/\Delta$ based on epidemiological plausibility, where Δ is the maximum distance between CTs in LA county. We use 60 design points per CT and select 150 knots throughout the study region using the space filling design described in Johnson et al. (1990) and implemented in the R package FIELDS.

We only display results from the spatio-temporal models for the total population. In

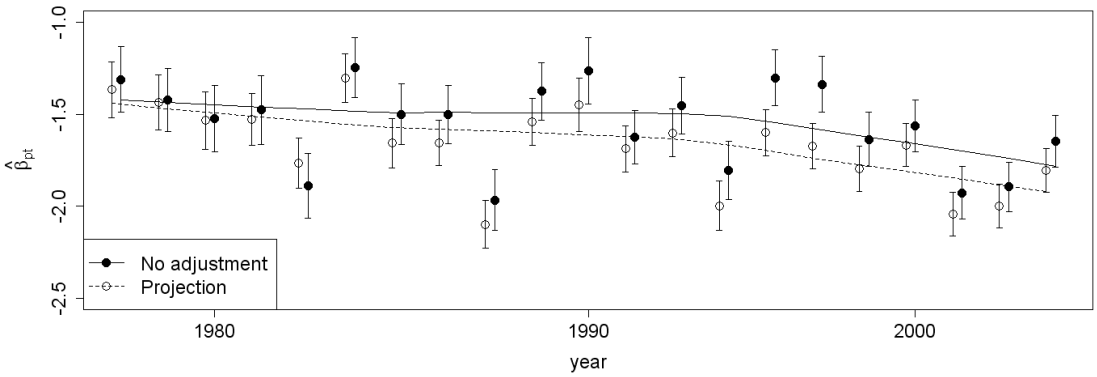


Figure 5.4: Estimates of $\hat{\beta}_{pt}$ and corresponding 50% confidence intervals for the total population, using the spatial model with and without the spatial confounding projection.

the black population, estimates of the variance components $\{\sigma_t\}$ were consistently zero, suggesting no residual spatial variability exists in the black population.

Results from the total population are plotted in Figure 5.4. Examining the plot, we see that the socioeconomic gradient is attenuated in the unadjusted model that ignores the impact of spatial confounding, compared to the model with the projected basis functions.

Next, we assume a linear trend in breast cancer incidence and in the socioeconomic gradient over time,

$$log(\mu_{it}) = \beta_0 + \beta^T t + \beta^P x_{it} + \beta^{PT} x_{it} t + \boldsymbol{Z_{it}(s_{it})u_t}.$$

Ignoring spatial confounding, we estimate $\hat{\beta}_{PT} = -0.014$, and $se(\hat{\beta}_{PT}) = 0.009$. Testing the hypothesis that $\hat{\beta}^{PT} = 0$, the p-value is 0.11, and we do not observe a statistically significant change in the socioeconomic gradient over time. Including the spatial projection in Section 5.5.3, we estimate $\hat{\beta}_{PT} = -0.023$, and $se(\hat{\beta}_{PT}) = 0.007$. Now, we do observe a statistically significant increase in the socioeconomic gradient over time ($p = 0.001$). Without projecting, spatial confounding could mask the increasing gradient in socioeconomic status in LA county. Results from the spatial confounding model are similar to the results from the overdispersed Poisson model.

5.7 Discussion

In this paper, we address common issues in large disease mapping applications, including missing population counts at intercensal years, temporal boundary misalignment, and spatial confounding. We assess the impact of uncertainty in intercensal population counts on estimating data associations between area-level indicators and disease incidence. When the intercensal interpolation model induces correlation between the indicators and errors in the expected case count, bias can be substantial. By using model stacking, we reduce the prediction error in our model and thereby should lower the covariance between the ABSM and these errors. In the future, we plan to explore accounting for model uncertainty using Bayesian model averaging (Hoeting et al., 1999).

We also propose a model for temporal boundary misalignment within a geostatistical framework, generalizing the model presented in Hund et al. (2012). Within this framework, it is simple to implement the spatial confounding projection to avoid collinearity between fixed effects and spatial random effects (Hodges and Reich, 2010).

Analyzing data from women in LA county between 1980 and 2000, we find that the socioeconomic gradient in breast cancer incidence does not appear to be decreasing over time. Rather, in the total population, the gradient appears to increase over the twenty-year period. If we ignored spatial confounding in our analyses, we would have markedly underestimated the increase in the socioeconomic gradient in breast cancer incidence.

Chen et al. (2008) emphasize that it may not be appropriate to assume a common spatial effect across racial/ethnic groups due to patterns of racial/ethnic segregation. In future analyses, we hope to expand our analysis to other race/ethnicity groups.

5.7.1 Incorporating uncertainty in intercensal counts

In future analyses, we aim to incorporate uncertainty in the intercensal population counts into our model. When interpolating intercensal tract counts, there are two primary

sources of uncertainty: (1) uncertainty associated with the kriging predictions during the boundary re-normalization, and (2) uncertainty associated with the choice model for the intercensal counts.

Using the population apportionment model in Section 5.3.2, Best and Wakefield (1999) suggest modelling intercensal population counts using a hierarchical Bayesian framework:

$$N_{1ta}, ..., N_{M_nta}|N_t \sim \text{Multinomial}(N_{ta}, p_{1ta}, ..., p_{nta}),$$
$$p_{1ta}, ..., p_{M_nta} \sim \text{Dirichlet}(s_{1ta}\hat{p}_{1ta}, ..., s_{nta}\hat{p}_{nta})$$

where $\hat{p}_{it} = N_{it}/\sum_i N_{it}$, and n is the total number of census tracts at the intercensal years.

The parameters $\{s_{ita}\}$ control the variance of the Dirichlet prior and consequently control the amount of uncertainty in the intercensal count estimates. Best and Wakefield (1999) use migration data from the previous year to calibrate their model and estimate the $\{s_{ita}\}$ parameters.

The Best and Wakefield (1999) model for incorporating uncertainty has several limitations for our application. First, BW introduce sampling error by modelling $N_{ita}|N_{ta}$ using the multinomial model. In our application, we do not actually have any "sampling error", because we are dealing with a census of the population. The error in the intercensal estimates is driven by model misspecification, not sampling. Additionally, we are interested in assessing the change in an association over time. The BW model conditions on the county-level total at each time point and consequently interpolates intercensal counts independently at each time point. We anticipate that counts within a tract at consecutive time points will be highly correlated. When examining changes in the association between disease incidence and an indicator (SES) over time, we lose efficiency by ignoring autocorrelation in the population denominators. We prefer a method for incorporating uncertainty that preserves the temporal correlation in census tract counts.

References

AMERICAN CANCER SOCIETY (2010). Statistics: 2010. (http://www.cancer.org) [accessed: October 1, 2010].

ANKER, M., BLACK, R., COLDHAM, C., KALTER, H., QUIGLEY, M., ROSS, D. and SNOW, R. (1999). *A standard verbal autopsy method for investigating causes of death in infants and children*. World Health Organization Geneva.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall.

BANERJEE, S., GELFAND, A., FINLEY, A. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 825–848.

BÄRNIGHAUSEN, T., BOR, J., WANDIRA-KAZIBWE, S. and CANNING, D. (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology* **22** 27.

BECKER, S., THORNTON, J. and HOLDER, W. (1993). Infant and child mortality estimates in two counties of Liberia: 1984. *International journal of epidemiology* **22** S42–S49.

BESAG, J., YORK, J. and MOLLIE, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43** 1–20.

BEST, N., RICHARDSON, S. and THOMPSON, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* **14** 35–59.

BEST, N. and WAKEFIELD, J. (1999). Accounting for inaccuracies in population counts and case registration in cancer mapping studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162** 363–382.

BEST, N. G., ICKSTADT, K. and WOLPERT, R. L. (2000). Spatial poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association* **95** 1076–1088.

BILDER, C. and TEBBS, J. (2009). Bias, efficiency, and agreement for group-testing regression models. *Journal of statistical computation and simulation* **79** 67–80.

BILDER, C., TEBBS, J. and CHEN, P. (2010). Informative retesting. *Journal of the American Statistical Association* **105** 942–955.

BILUKHA, O. (2008). Old and new cluster designs in emergency field surveys: in search of a one-fits-all solution. *Emerging Themes in Epidemiology* **5**.

BILUKHA, O. and BLANTON, C. (2008). Interpreting results of cluster surveys in emergency settings: is the LQAS test the best option. *Emerging Themes in Epidemiology* **5** 25.

BINKIN, N., SULLIVAN, K., STAEHLING, N. and NIEBURG, P. (1992). Rapid nutrition surveys: How many clusters are enough? *Disasters* **16** 97.

BOYLE, P. and PARKIN, D. (1991). Statistical methods for registries. *Cancer Registration Principles and Methods. Lyon: IARC* 126–58.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88** 9–25.

BURROWS, P. (1987). Improved estimation of pathogen transmission rates by group testing. *Phytopathology* **77** 363–365.

CASTRO, A. and FARMER, P. (2005). Understanding and addressing AIDS-related stigma: From anthropological theory to clinical practice in Haiti. *American Journal of Public Health* **95** 53–9.

CENTRAL STATISTICAL AGENCY and ORC MACRO (2006). Ethiopia Demographic and Health Survey 2005.

CHEN, J. T., COULL, B. A., WATERMAN, P. D., SCHWARTZ, J. and KRIEGER, N. (2008). Methodologic implications of social inequalities for analyzing health disparities in large spatiotemporal data sets: An example using breast cancer incidence data (Northern and Southern California, 1988-2002). *Statistics in Medicine* **27** 3957–3983.

CHEN, P., TEBBS, J. and BILDER, C. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65** 1270–1278.

COLOSIMO, B. and DEL CASTILLO, E. (2006). *Bayesian process monitoring, control and optimization*. Chapman & Hall/CRC.

COULL, B. A., RUPPERT, D. and WAND, M. P. (2001). Simple incorporation of interactions into additive models. *Biometrics* **57** 539–545.

COX, L. and ZAYATZ, L. (1995). An agenda for research on statistical disclosure limitation. *Journal of Official Statistics* **11** 205–20.

CRAINICEANU, C., RUPPERT, D. and WAND, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software* **14** 1–24.

CURTIS, S. and SUTHERLAND, E. (2004). Measuring sexual behaviour in the era of HIV/AIDS: the experience of Demographic and Health Surveys and similar enquiries. *Sexually Transmitted Infections* **80** ii22–ii27.

DE WALQUE, D. (2007). Sero-discordant couples in five african countries: Implications for prevention strategies. *Population Development Review* **33** 501–523.

DEITCHLER, M., DECONINCK, H. and BERGERON, G. (2008). Precision, time, and cost: a comparison of three sampling designs in an emergency setting. *Emerging Themes in Epidemiology* **5** 6.

DEITCHLER, M., VALADEZ, J., EGGE, K., FERNANDEZ, S. and HENNIGAN, M. (2007). A field test of three LQAS designs to assess the prevalence of acute malnutrition. *International journal of epidemiology* **36** 858.

DIAZ, T., DE COCK, K., BROWN, T., GHYS, P. and BOERMA, J. (2005). New strategies for HIV surveillance in resource-constrained settings: an overview. *AIDS* **16** S1–S8.

DIXON, P., ELLISON, A. and GOTELLI, N. (2005). Improving the precision of estimates of the frequency of rare events. *Ecology* **86** 1114–1123.

DODGE, H. and ROMIG, H. (1929). A method of sampling inspection. *The Bell System Technical Journal* **8** 398.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*. U.S. Office of Management and Budget, Washington, DC.

FENN, B., MORRIS, S. and FROST, C. (2004). Do childhood growth indicators in developing countries cluster? Implications for intervention strategies. *Public health nutrition* **7** 829–834.

GARCIA-CALLEJA, J., GOUWS, E. and GHYS, P. (2006). National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates. *Sexually Transmitted Infections* **82** iii64–iii70.

GASTWIRTH, J. and HAMMICK, P. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. *Journal of Statistical Planning and Inference* **22** 15–27.

GOTWAY, C. and YOUNG, L. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97** 632–648.

GOTWAY, C. and YOUNG, L. (2007). A geostatistical approach to linking geographically

aggregated data from different sources. *Journal of Computational and Graphical Statistics* **16** 115–135.

GOUWS, E., MISHRA, V. and FOWLER, T. (2008). Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data. *Sexually Transmitted Infections* **84** i17–i23.

GREENLAND, K., RONDY, M., CHEVEZ, A., SADOZAI, N., GASASIRA, A., ABANIDA, E., PATE, M., RONVEAUX, O., OKAYASU, H., PEDALINO, B. ET AL. (2011). Clustered lot quality assurance sampling: a pragmatic tool for timely assessment of vaccination coverage. *Tropical Medicine & International Health* **16** 863–868.

HARVILLE, D. (2008). *Matrix algebra from a statistician's perspective*. Springer Verlag.

HEDT-GAUTHIER, B., MITSUNAGA, T., HUND, L., OLIVES, C. and PAGANO, M. (2012). An approach for incorporating clustering effects into the design of lot quality assurance sampling.

HEPWORTH, G. and WATSON, R. (2009). Debiased estimation of proportions in group testing. *Journal of the Royal Society of Statistics (Series C)* **58** 105–121.

HODGES, J. and REICH, B. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* **64** 325–334.

HOETING, J., MADIGAN, D., RAFTERY, A. and VOLINSKY, C. (1999). Bayesian model averaging: a tutorial. *Statistical science* 382–401.

HUND, L., CHEN, J., KRIEGER, N. and COULL, B. (2012). A geostatistical approach to large-scale disease mapping with temporal misalignment.

JOHNSON, M., MOORE, L. and YLVISAKER, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26** 131.148.

KAMMANN, E. and WAND, M. (2003). Geoadditive models. *Applied Statistics* **52** 1–18.

KATZ, J. (1995). Sample-size implications for population-based cluster surveys of nutritional status. *The American journal of clinical nutrition* **61** 155–160.

KELSALL, J. and WAKEFIELD, J. (2002). Modeling spatial variation in disease risk: A geostatistical approach. *Journal of the American Statistical Association* **97** 692–701.

KERRY, S. and MARTIN BLAND, J. (2001). Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in medicine* **20** 377–390.

KRIEGER, N., CHEN, J., WATERMAN, P., SOOBADER, M., SUBRAMANIAN, S. and CARSON, R. (2002). Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? *American journal of epidemiology* **156** 471–482.

KRIEGER, N., CHEN, J. T., WATERMAN, P. D., REHKOPF, D. H., YIN, R. and COULL, B. A. (2006). Race/ethnicity and changing U.S. socioeconomic gradients in breast cancer incidence: California and Massachusetts, 1978-2002. *Cancer Causes and Control* **17** 217–226.

KRISHNAN, S. and JESANI, A. (2009). Unlinked anonymous HIV testing in population-based surveys in India. *Indian Journal of Medical Ethics* **6** 182–184.

LEE, D.-J. and DURBAN, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics and Data Analysis* **53** 2968–2979.

LEMESHOW, S. and ROBINSON, D. (1985). Surveys to measure programme coverage and impact: a review of the methodology used by the expanded programme on immunization. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales* **38** 65.

LITTLE, R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* **54** 139–157.

LITTLE, R. (1988). Missing data adjustments in large surveys. *Journal of Business and Economic Statistics* **6** 287–296.

LITTLE, R. and RUBIN, D. (2002). *Statistical Analysis with Missing Data*. 2nd ed. Wiley and Sons, New York.

LITTLE, R. and VARTIVARIAN, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine* **22** 1589–1599.

LITVAK, E., TU, X. and PAGANO, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association* **89** 424–434.

LOHR, S. (1999). *Sampling: Design and Analysis*. Brooks/Cole.

MARTIN-HERZ, S., SHETTY, A., BASSETT, M., LEY, C., MHAZO, M., MOYO, S., HERZ, A. and KATZENSTEIN, D. (2006). Perceived risks and benefits of HIV testing, and predictors of acceptance of HIV counselling and testing among pregnant women in Zimbabwe. *International Journal of Sexually Transmitted Diseases and AIDS* **17** 835–841.

MCMAHAN, C., TEBBS, J. and BILDER, C. (2011). Informative dorfman screening. *Biometrics* .

MILLER, R. (1974). The jackknife-a review. *Biometrika* **61** 1–15.

MISHRA, V., BARRERE, B., HONG, R. and KHAN, S. (2008). Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sexually Transmitted Infections* **84** i63–i70.

MUGGLIN, A. S., CARLIN, B. P. and GELFAND, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association* **95** 877–887.

MULLER, H.-G., STADTMULLER, U. and TABNAK, F. (1997). Spatial smoothing of geographically aggregated data, with application to construction of incidence maps. *Journal of the American Statistical Association* **92** 61–71.

MYATT, M. and BENNETT, D. (2008). A novel sequential sampling technique for the surveillance of transmitted HIV drug resistance by cross-sectional survey for use in low resource settings. *Antiviral therapy* **13** 37.

MYATT, M., DUFFIELD, A., SEAL, A. and PASTEUR, F. (2009). The effect of body shape on weight-for-height and mid-upper arm circumference based case definitions of acute malnutrition in Ethiopian children. *Annals of Human Biology* **36** 5–20.

NATIONAL POPULATION COMMISSION and ICF MACRO (2009). *Nigeria Demographic and Health Survey 2008*. National Population Commission and ICF Macro, Abuja, Nigeria.

NATIONAL STATISTICAL OFFICE AND ORC MACRO (2005). *Malawi Demographic and Health Survey 2004*. Calverton, Maryland.

NATIONAL STATISTICS OFFICE and ICF MACRO (2009). *Philippines National Demographic and Health Survey 2008*. National Statistics Office and ICF Macro, Calverton, Maryland.

OBERMEYER, C. and OSBORN, M. (2007). The utilization of testing and counseling for HIV: A review of the social and behavioral evidence. *American Journal of Public Health* **97** 1–13.

OLIVES, C. and PAGANO, M. (2010). Bayes-LQAS: classifying the prevalence of global acute malnutrition. *Emerging Themes in Epidemiology* **7** 3.

OLIVES, C., PAGANO, M., DEITCHLER, M., HEDT, B., EGGE, K. and VALADEZ, J. (2009). Cluster designs to assess the prevalence of acute malnutrition by lot quality assurance sampling: a validation study by computer simulation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172** 495–510.

PEZZOLI, L., PINEDA, S., HALKYER, P., CRESPO, G., ANDREWS, N. and RONVEAUX, O. (2009). Cluster-sample surveys and lot quality assurance sampling to evaluate yellow fever immunisation coverage following a national campaign, Bolivia, 2007. *Tropical Medicine & International Health* **14** 355–361.

PHIPPS, A., CLARKE, C., EREMAN, R. ET AL. (2005). Impact of intercensal population projection and error of closure on breast cancer surveillance: Examples from 10 California counties. *Breast Cancer Res* **7** R655–660.

QUENOUILLE, M. (1956). Notes on bias in estimation. *Biometrika* **43** 353–60.

QUINN, T., THOMAS, C., BROOKMEYER, R., KLINE, R., SHEPHERD, M., PARANJAPE, R., MEHENDALE, S., GADKARI, D. and BOLLINGER, R. (2000). Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence. *AIDS* **14** 2751–2757.

RAO, J. and SCOTT, A. (1992). A simple method for the analysis of clustered binary data. *Biometrics* **48** 577–585.

REICH, B., HODGES, J. and ZADNIK, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62** 1197–1206.

RENIERS, G., ARAYA, T., BERHANE, Y., DAVEY, G. and SANDERS, E. (2009). Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. *BMC Public Health* **9** 163–172.

RENIERS, G. and EATON, J. (2009). Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS* **23** 621–629.

RIDOUT, M., DEMÉTRIO, C. and FIRTH, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55** 137–148.

ROBERTSON, S. and VALADEZ, J. (2006). Global review of health care surveys using lot quality assurance sampling (LQAS), 1984-2004. *Social Science & Medicine* **63** 1648–1660.

ROTHENBERG, R., LOBANOV, A., SINGH, K. and STROH JR, G. (1985). Observations on the application of EPI cluster survey methods for estimating disease incidence. *Bulletin of the World Health Organization* **63** 93.

RUPPERT, D., WAND, M. and CARROLL, R. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

SANDIFORD, P. (1993). Lot quality assurance sampling for monitoring immunization programmes: cost-efficient or quick and dirty? *Health policy and planning* **8** 217.

SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

SHISANA, O., SIMBAYI, L., PARKER, W., ZUMA, K., BHANA, A., CONNOLLY, C., JOOSTE, S. and PILLAY, V. (2005). *South African HIV Prevalence, HIV Incidence, Behaviour and Communication Study*. 2nd ed. HSRC Press, Cape Town.

SMITH, S. and SHAHIDULLAH, M. (1995). An evaluation of population projection errors for census tracts. *Journal of the American Statistical Association* **90** 64–71.

SOKAL, D., IMBOUA-BOGUI, G., SOGA, G., EMMOU, C. and JONES, T. (1988). Mortality from neonatal tetanus in rural Côte d'Ivoire. *Bulletin of the World Health Organization* **66** 69.

SOWER, V., MOTWANI, J. and SAVOIE, M. (1993). Are acceptance sampling and SPC complementary or incompatible? *Quality Progress* **26** 85–85.

SPIEGEL, P. (2007). Who should be undertaking population-based surveys in humanitarian emergencies? *Emerging Themes in Epidemiology* **4** 12.

STROH, G. and BIRMINGHAM, M. (2002). Protocol for assessing neonatal tetanus mortality in the community using a combination of cluster and lot quality assurance sampling. URL `www.who.int/vaccines-documents/`

TANSER, F., HOSEGOOD, V., BÄRNIGHAUSEN, T., HERBST, K., NYIRENDA, M., MUHWAVA, W., NEWELL, C., VILJOEN, J., MUTEVEDZI, T. and NEWELL, M. (2008). Cohort profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *International Journal of Epidemiology* **37** 956–962.

TOBLER, W. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 519–530.

TOURANGEAU, R. and YAN, T. (2007). Sensitive questions in surveys. *Psychological Bulletin* **133** 859–883.

TU, X., LITVAK, E. and PAGANO, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika* **82** 287–97.

US CENSUS BUREAU (1994). Geographical areas reference manual. Tech. rep.

VALADEZ, J. (1991). *Assessing child survival programs in developing countries: testing lot quality assurance sampling; Assessing child survival programs in developing countries: testing lot quality assurance sampling*. Harvard School of Public Health. Department of Population and International Health.

VANSTEELANDT, S., GOETGHEBEUR, E. and VERSTRAETEN, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **54** 1126–33.

VERMUND, S. H. and WILSON, C. M. (2002). Barriers to HIV testing-where next? *Lancet* **360** 1186–7.

WAGER, C., COULL, B. and LANGE, N. (2004). Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging. *Journal of the Royal Statistical Society, Series B* **66** 429–446.

WAKEFIELD, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* **8** 158–183.

WAND, M. (2003). Smoothing and mixed models. *Computational Statistics* **18** 223–249.

WOLPERT, D. (1992). Stacked generalization. *Neural networks* **5** 241–259.

WORLD HEALTH ORGANIZATION (1999). *Management of severe malnutrition: a manual for physicians and other senior health workers*. World Health Organization.

WORLD HEALTH ORGANIZATION (2000). *The management of nutrition in major emergencies*. World Health Organization.

XIE, M. (2001). Regression analysis of group testing samples. *Statistics in Medicine* **20** 1957–1969.

YOUSRY, M., STURM, G., FELTZ, C. and NOOROSSANA, R. (1991). Process monitoring in real time: Empirical bayes approach: discrete case. *Quality and reliability engineering international* **7** 123–132.

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99** 250–261.

ZHU, L., CARLIN, B., ENGLISH, P. and SCALF, R. (2000). Hierarchical modeling of spatio-temporally misaligned data: relating traffic density to pediatric asthma hospitalizations. *Environmetrics* **11** 43–61.